UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS
MATEMÁTICAS

FACULTAD DE CIENCIAS

ESTIMACION DE VARIANZAS PARA ESTIMADORES NO LINEALES.

T E S I S

QUE PARA OBTENER EL GRADO ACADEMICO DE

MAESTRO EN CIENCIAS (MATEMÁTICAS)

P R E S E N T A :

RIGOBERTO REAL MIRANDA

DIRECTORA DE TESIS: DRA. GUILLERMINA ESLAVA GÓMEZ

MÉXICO, D. F. OCTUBRE 2006

Agradecimientos

A mis **padres**, quienes siempre me han apoyado en mi carrera profesional, así como a hermanos y familiares.

A mis **amigos**, que siempre estuvieron para escucharme cuando lo necesitaba y con quienes he compartido situaciones tanto buenas como no tan buenas.

A mis **compañeros Banxico**, y en especial a Graciela Ruíz, quienes ejercieron presión y me apoyaron para terminar este trabajo.

A mis **compañeras de jazz**, quienes siempre me regalaban una sonrisa de aliento en todo momento.

A los honorables miembros del jurado:

Dra. Guillermina Eslava Gómez

Dra. Ruth Selene Fuentes García

Dr. Ignacio Méndez Ramírez

M. en C. Leticia Eugenia Gracia-Medrano Valdelamar

Dr. Martín Romero Martínez

por su acertada orientación en la realización del mismo.

Un agradecimiento especial a la **Dra. Guillermina Eslava** por haberme apoyado en la realización de este trabajo, así como a todas aquellas personas que directa o indirectamente participaron en el mismo.

A la **Universidad Nacional Autónoma de México**, al **IIMAS** y a todos sus profesores.

Un agradecimiento sincero al **Lic. Juan Francisco Sánchez Martínez** por estar a mi lado en esta última etapa del tan largo y complicado camino y por haberme enseñado una nueva filosofía de la vida.

Rigoberto Real Miranda

México, D. F.
Octubre de 2006

Índice general

Índice de tablas

Índice de figuras

Resumen

El cálculo de la varianza de estimadores no lineales es un problema en ocasiones difícil de resolver debido a la forma analítica que presentan tales estimadores, por ejemplo los estimadores de razón, además de las características del diseño muestral con la cual fue obtenida la información.

En este trabajo, se revisan la expresiones de la estimación puntual de tres parámetros no-lineales de interés que son: la Mediana, el índice de Gini y la estimación de un índice propuesto por Kish, conocido como el índice de Kish. Además de la estimación puntual de estos parámetros, se presenta una estimación de su varianza, considerando el diseño muestral complejo con el cual fue obtenida la información. Para el caso del índice de Gini se presentan dos métodos adicionales a la estimación de su varianza. La estimación de la varianza del índice de Kish se hizo bajo el supuesto de que las muestras utilizadas no eran independientes, esto es, que eran del tipo panel, por lo cual se incluyó el término de la covarianza en tal estimación.

Para ilustrar el procedimiento de la estimación puntual y de la varianza de estos parámetros, se utilizaron las bases de datos muestrales del XII Censo de Población y Vivienda 2000 y de la Encuesta Nacional de Ingresos y Gastos en Hogares, en sus ediciones 2000, 2002 y 2004 obtenidas por el INEGI.

Los recursos computacionales usados para obtener las estimaciones de tales parámetros fueron: el paquete estadístico **SUDAAN 7.5.6**, que sirvió en la estimación de la Mediana del Ingreso Promedio en Hogares y de su varianza, el paquete estadístico **S**-Plus para programar la rutina de cálculo de la estimación del índice de Gini del Ingreso en Hogares y, finalmente, el paquete estadístico **R**, en el cual se programó la rutina para la estimación del índice de Kish y de su varianza para los años 2002 y 2004.

Los paquetes estadísticos **S**-Plus y **R** se usaron debido a que permitían el manejo de bases de datos y la programación de rutinas de cálculo complejas para la estimación de los parámetros mencionados, cosa que paquetes como **SUDAAN 7.5.6** y **SPSS** no permitieron hacer.

Notación

a. Poblacional, Variables.

$U = \{u_1, u_2, \ldots, u_k, \ldots, u_N\}$	Población (=conjunto de N elementos)
U_h	Estrato h $(h = 1, \ldots, H)$
y_k	Valor de la variable de estudio y para el elemento k
x_{jk}	Valor de la variable auxiliar x_j para el elemento k
$t_y = t_{yU} = \sum_U y_k$	Total de la Población de la variable y
$\bar{y}_U = \frac{t_y}{N}$	Media de la Población de la variable$y$
$S^2_{yU} = \frac{\sum_U (y_k - \bar{y}_U)^2}{N-1}$	Varianza de la Población de la variable y
$S^2_{yzU} = \frac{\sum_U (y_k - \bar{y}_U)(z_k - \bar{z}_U)}{N-1}$	Covarianza de la Población de las variables y y z
$R = t_{yU}/t_{zU}$	Razón de la Población entre los totales (medias) de las variables y y z
M	Mediana de la población.
R_N	Coeficiente de Gini

b. Muestral.

s	Muestra (subconjunto de U)
n_s	Tamaño muestral (número de elementos en s)
$\bar{y}_s = \sum_s y_k/n_s$	Media muestral de y
$S^2_{ys} = \frac{\sum_s (y_k - \bar{y}_s)^2}{n_s - 1}$	Varianza muestral de y
$p(s)$	Probabilidad de selección de la muestra s
$p(\cdot)$	Diseño muestral
π_k	Probabilidad de inclusión (primer orden) del elemento k
π_{kl}	Probabilidad de inclusión (segundo orden) de los elementos k y l

c. Estimadores.

$\hat{\theta}$	Estimador de un parámetro general θ
$V\left(\hat{\theta}\right)$	Varianza del estimador $\hat{\theta}$
$\widehat{V}\left(\hat{\theta}\right)$	Estimador de $V\left(\hat{\theta}\right)$
$\hat{t}_{y\pi}$	Estimador Horvitz-Thompson, π del total poblacional $t_y = \sum_U y_k$
$\hat{N} = \sum_s \frac{1}{\pi_k}$	Estimador π del total poblacional
$\hat{M}$	Estimador de la Mediana
$\hat{R}_N$	Estimador Coeficiente de Gini
$\hat{r} = \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}$	Estimador de razón de la variable y respecto a x.
$\hat{R}_1 = \frac{\hat{r}_1}{\hat{r}_0}$	Estimador de razón doble, o relativa, del periodo 1 respecto al periodo 0.
$\widehat{C}\left(\hat{r}_0, \hat{r}_1\right)$	Estimador de la covarianza de los estimadores de razón entre periodos 0 y 1

Introducción

Este trabajo está enfocado en la estimación de parámetros de interés práctico que no son muy discutidos en la literatura usual del muestreo de encuestas, tales como Cochran (1976), Särndal (1992), debido a la complejidad que muestran en su proceso: Mediana, Indice de Gini, Indices basados en estimadores de Razón. Además de esto, se presenta la estimación de sus varianzas, la cual es importante obtenerla ya que permite conocer la variabilidad que presentan tales estimadores.

El primer tema que se presenta es la metodología de la estimación de la Mediana de una Población cuando se considera un diseño muestral complejo, que a simple vista puede parecer un tema sencillo, pero que tratado más a detalle uno se encuentra con la dificultad de que está basado en una muestra s, la cual es probabilística y que a cada elemento de la población $y_k,\ k \in U$, le asigna una probabilidad de ser seleccionada mayor a cero. Por lo que cada elemento tiene un peso diferente, razón por la cual se complica el problema de estimación. De igual forma, no sólo está el problema de obtener un estimador de la mediana sino también el de presentar un intervalo de confianza tal estimación.

En este caso, se presenta un método inicialmente propuesto por Woodruff (1952), quién, en forma gráfica, obtuvo los intervalos de confianza para la estimación de la mediana, sin tener tanto rigor estadístico y matemático. Posteriormente, Francisco y Fuller (1991), basándose en esta idea, desarrollaron con más rigor estadístico las expresiones explícitas de los límites de confianza, llegando a los mismos planteados por Woodruff, obteniendo también la justificación matemática de tales intervalos en el teorema que presentan en su artículo.

Otro tema a tratar en este trabajo es la estimación del coeficiente de Gini, el cual es un índice que mide la distribución de una característica, e.g. del ingreso, consumo, o riqueza, en una población dada. Los métodos propuestos para la estimación de este coeficiente son diversos, debido a la complejidad que representa su cálculo, ya que se requiere que la información esté ordenada y acumulada por la variable de interés. Aunado a esto, en el contexto de estimación con una muestra que proviene de un diseño complejo, los cálculos no siempre se facilitan. Aquí se presentan dos opciones del estimador de varianza para el estimador del índice de Gini, uno propuesto por Ogwang (2000) que se basa en la técnica jackknife y no considera diseño muestral complejo en su estimación; el otro, fue propuesto por Sandström et al. (1988), quienes presentan un estimador al coeficiente de Gini considerando un diseño muestral complejo, además de que presentan una expresión explícita para estimar la varianza del coeficiente de Gini propuesto cuando las probabilidades de selección de cada elemento son diferentes.

Finalmente, se revisa el tema de estimación de varianzas para índices. El método presentado para la obtención de índices es el propuesto por Kish (1957) y que es tratado como un estimador de razón. De igual forma, en ese mismo artículo, Kish propone un estimador de la varianza para índices, que se reduce a la obtención de estimadores de varianza de un estimador de razón. Este tema puede ser revisado a detalle en Särndal et al. (1992), quienes presentan la forma de obtener estimadores de razón cuando se considera un diseño muestral complejo.

Para ejemplificar las estimaciones mencionadas a casos prácticos, se utilizaron varias fuentes de información. Una fue la del XII Censo General de Población y Vivienda del 2000, que sirvió para la obtención de la Mediana del Ingreso por Hogares (capítulo 1) y para la estimación del Coeficiente de Gini (capítulo 2), la otra fue la Encuesta Nacional de Ingresos y Gastos en Hogares (ENIGH), en sus ediciones 2000, 2002 y 2004, que sirvió de base en la construcción de índices, y de sus varianzas, del gasto promedio total en hogares (capítulo 3). Además de la información que publica el Banco de México referente al Indice Nacional de Precios al Consumidor, INPC (véase [16]).

La estimación de los parámetros aquí presentados se realizaron auxiliándose de varios paquetes estadísticos que permiten considerar un diseño muestral complejo (capítulo A). La estimación de la Mediana y su varianza se obtuvieron con el paquete estadístico **SUDAAN 7.5.6**. La estimación de índices y su varianza fue programada en el paquete **R 2.2.0**, con ayuda de la subrutina **survey**, que permite sólo obtener estimación de totales y de razón a un nivel muy básico. Para la estimación del índice de Gini y su varianza, los procedimientos fueron programados bajo el ambiente de **S**-Plus y **R**, debido a que no hay paquete que realice los cálculos de las expresiones aquí presentadas para este último parámetro.

Capítulo 1

Estimación de la Mediana Poblacional

Introducción

Cuando se desea obtener la estimación de la mediana de una población, la mayoría de las veces se desconoce la función de distribución de la que proviene. El método descrito a continuación no considera el supuesto de una distribución de la variable de interés. Se requiere sólo que la muestra obtenida sea probabilística, i.e., que la probabilidad de cada elemento de la población sea conocida.

1.1. Método de Estimación

La Mediana de una población con respecto a una variable y es un valor M que divide a la población a la mitad. De esta manera, la mitad de los elementos de la población tienen valores menores a M, mientras que la otra mitad tienen valores más grandes a M.

En otras palabras, dada una población $y_1, \ldots, y_N$; M corresponde al valor para el cual la mitad de las observaciones son menores y la mitad son mayores.

Sea $F(y)$ la función de distribución acumulativa empírica

$$F(y) = \frac{1}{N}(\#A_y)$$

donde A_y es el conjunto de los elementos poblacionales cuyo y_k es menor o igual a y, esto es, $A_y = \{k : k \in U, \text{ y } y_k \leq y\}$ y $\#A_y$ denota el número de elementos, o la cardinalidad, en el conjunto A_y.

La mediana de la población M se define entonces como

$$M = F^{-1}(0.5)$$

donde $F^{-1}(\cdot)$ es la función inversa de $F(\cdot)$. La igualdad estricta se obtiene cuanto la función de distribución, $F(\cdot)$, es continua.

El problema ahora es estimar la mediana poblacional M, usando datos y_k para $k \in s$, donde s es una muestra probabilística. A continuación se describe el método de estimación con más detalle, incluyendo la obtención de un intervalo de confianza propuesto por Woodruff (1952).

Se selecciona una muestra s de la población mediante un diseño muestral $p(\cdot)$ con probabilidades de inclusión π_k, π_{kl}. Escríbase a $F(y)$ como

$$F(y) = \frac{1}{N}\sum_U z_{ky} = \frac{t_{z_y}}{N} = \bar{z}_{yU} \qquad z_{ky} = \begin{cases} 1 & \text{si } \; y_k \leq y \\ 0 & \text{si } \; y_k > y \end{cases} \tag{1.1}$$

de esta manera, la estimación de (1.1) está dada por

$$\hat{F}(y) = \tilde{z}_{y,s} = \frac{\hat{t}_{z_y,\pi}}{\hat{N}} = \frac{\sum_s \frac{z_{ky}}{\pi_k}}{\sum_s \frac{1}{\pi_k}}. \tag{1.2}$$

La función de distribución estimada es una función no-decreciente, con saltos de cero a uno, debido a que la función (1.1) es una función discreta. Cuando en (1.2) se divide por $\hat{N}$ en lugar de N (si fuera conocida) tiene la ventaja de que $\hat{F}(y)$ alcanza el valor de la unidad a medida que y se incrementa, la cual es una propiedad deseable de la función de distribución, en cambio, si $F(y)$ fuera estimada por $N^{-1}\sum_s z_{ky}/\pi_k$, el último valor, $N^{-1}\sum_s 1/\pi_k$, no necesariamente es la unidad. El número de valores que toma la variable y en la función de distribución es criterio del analista, ya que no hay un número fijo establecido.

Una vez encontrada la estimación de la función de distribución, la estimación de la mediana es

$$\hat{M} = \hat{F}^{-1}(0.5) \tag{1.3}$$

donde $\hat{F}^{-1}$ es la función inversa de $\hat{F}(y)$ dada por (1.2).

Una forma explícita a la estimación dada en (1.3) es el estimador Horvitz-Thompson dado por

$$\hat{M} = \frac{\sum_{k=1}^{n} \frac{W_{k,L}}{\pi_k}}{\sum_{k=1}^{n} \frac{1}{\pi_k}}, \quad \text{donde } W_{k,L} = \begin{cases} 1 & si \; y_k \leq L \\ 0 & si \; y_k > L \end{cases} \tag{1.4}$$

donde L es el límite superior de la variable de estudio que determina los valores que estarán por debajo de la mediana.

Sólo falta encontrar un intervalo de confianza aproximado para M. Para dos constantes cualquiera c_1 y c_2 y para cualquier valor de M,

$$\Pr\left\{c_1 \leq \hat{F}\left(M\right) \leq c_2\right\} \doteq \Pr\left\{F^{-1}\left(c_1\right) \leq M \leq F^{-1}\left(c_2\right)\right\}. \tag{1.5}$$

Si F y $\hat{F}$ fueran continuas y estrictamente crecientes, entonces (1.5) sería igualdad exacta; la aproximación se debe a los saltos de las funciones F y $\hat{F}$.

De (1.5) se sigue que si c_1 y c_2 son encontradas tales que

$$\Pr\left\{c_1 \leq \hat{F}\left(M\right) \leq c_2\right\} = 0.95$$

entonces el intervalo

$$\left[\hat{F}^{-1}\left(c_1\right),\ \hat{F}^{-1}\left(c_2\right)\right]$$

sería aproximadamente un intervalo de confianza para M al 95 %. Usualmente, no se pueden determinar las constantes c_1 y c_2, así que se requiere de aproximaciones adicionales.

En Särndal et al. (1992), cap. 5, se hace el supuesto de que $\hat{F}$ se distribuye aproximadamente normal alrededor de su valor esperado, el cual es $F\left(M\right) \doteq 0.5$. Este supuesto se justifica para muestras grandes. Por lo que las constantes c_1 y c_2, serían

$$c_1 = 0.5 - 1.96\left\{V\left[\hat{F}\left(M\right)\right]\right\}^{1/2}, \qquad c_2 = 0.5 + 1.96\left\{V\left[\hat{F}\left(M\right)\right]\right\}^{1/2} \tag{1.6}$$

Para la expresión de la estimación de la varianza de $V\left[\hat{F}\left(M\right)\right]$, véase Särndal et al. (1992), cap. 5.

Existen paquetes estadísticos que obtienen la estimación de la mediana y de su varianza (o error estándar). En este trabajo se utilizó el paquete **SUDAAN**, que calcula la varianza de la función de distribución acumulativa de acuerdo al método de linealización de series de Taylor. Los límites del intervalo de confianza y la varianza son obtenidos con las fórmulas presentadas en Binder (1991) y Francisco y Fuller (1991). Cabe resaltar que los límites de confianza presentados por Francisco y Fuller son los mismos que los que presentó Woodruff (1952), sólo que Woodruff mostró una idea más gráfica de cómo obtenerlos, sin presentar una expresión analítica general del mismo. Sólo expone dos resultados particulares. En uno de ellos obtiene la varianza de la mediana mediante un muestreo aleatorio simple, y en el otro mediante un muestreo estratificado y aleatorio simple. Francisco y Fuller, sin embargo, en su artículo presentan la expresión de un intervalo de confianza, obtenido mediante funciones de distribución secuenciales. En el análisis que muestran, dan la justificación matemática del procedimiento del intervalo de confianza propuesto por Woodruff.

Para ilustrar el método de estimación de la mediana, y del cálculo de su intervalo de confianza, se utilizó la variable de Ingreso Mensual Promedio en Hogares para información proveniente de la muestra del XII Censo de Población y Vivienda 2000. El esquema de muestreo que presenta el Censo es por conglomerados en una sola etapa, es decir, se seleccionaron áreas geográficas completas, ya fueran áreas geoestadísticas básicas (Ageb), manzanas o localidades rurales. La selección de espacios geográficos completos permite aplicar un solo cuestionario en las viviendas: el ampliado en las seleccionadas y el básico en el resto. De esta manera, el total de la población se obtiene de sumar la información proveniente de los cuestionarios básico y ampliado.

El tamaño de muestra considerado en la estimación permitió el desglose de la información por Entidad Federativa y a nivel Nacional. El esquema de muestreo que se usó en este trabajo es estratificado y por conglomerados, en una sola etapa.

En la tabla 1.1 se muestran los resultados de la estimación puntual y el intervalo al 95 % de confianza de la Mediana del Ingreso Mensual Promedio por Hogares a nivel nacional y por entidad federativa. En estos resultados, se puede apreciar que los límites superior e inferior de los intervalos de confianza no son simétricos respecto a la estimación de la mediana. Incluso, se puede apreciar que en casos como Chiapas y Chihuahua la estimación de la mediana es muy parecida al límite inferior del intervalo y está muy alejado del límite superior. Caso contrario ocurre con los estados de Hidalgo y Michoacán, la estimación de la mediana está muy cercana al límite superior del intervalo de confianza.

El cálculo de la estimación de la mediana es sencillo, la estimación de su varianza no tanto debido a que se debe conocer la función de distribución de la variable de estudio, y de ahí, conocer su inversa, la cual no es fácil de obtener debido a que los datos, al ser muestrales, presentan distribuciones discretas.

En el siguiente capítulo se revisa la estimación de un parámetro que sirve para medir la distribución de la riqueza en una población dada. Además, se obtiene la estimación de su varianza aplicando diseño muestral complejo y diseño aleatorio simple.

Tabla 1.1: Estimación Puntual y por Intervalo al 95 % de confianza de la Mediana del Ingreso Mensual Promedio por Hogares (IMPH), Nacional y por Entidad Federativa considerando Diseño Muestral Complejo. Método propuesto originalmente por Woodruff (1952) y demostrado por Francisco y Fuller (1991) y Binder (1991). Cálculos basados en la muestra del XII Censo de Población y Vivienda 2000.

ESTADO	**TAMAÑO MUESTRA**	**MEDIANA IMPH PESOS**	**ERROR EST.**	**LÍMITE INF.**	**SUP.**
AGS	17,698	3,856	80.47	3,681	3,997
BC	33,910	6,212	116.65	5,977	6,434
BCS	8,610	4,708	82.50	4,571	4,895
CAM	14,500	2,386	79.40	2,218	2,529
COAH	44,269	3,914	56.81	3,823	4,046
COL	11,848	3,343	73.46	3,229	3,517
CHIS	80,681	1,286	30.59	1,285	1,405
CHIH	64,017	4,286	26.60	4,285	4,390
DF	180,982	4,627	37.33	4,554	4,701
DGO	30,475	2,957	54.08	2,840	3,052
GTO	70,668	3,168	56.9	3,062	3,285
GRO	62,615	2,057	45.08	1,963	2,140
HGO	55,532	2,142	26.87	2,037	2,143
JAL	132,796	4,094	33.74	4,025	4,157
MEX	231,073	3,213	16.17	3,191	3,254
MICH	88,319	2,570	4.00	2,555	2,571
MOR	33,339	2,786	33.79	2,733	2,866
NAY	17,144	2,654	125.64	2,517	3,009
NLN	76,400	4,913	49.21	4,822	5,015
OAX	125,936	1,557	35.03	1,505	1,642
PUE	115,994	2,270	48.14	2,231	2,420
QRO	25,375	3,598	114.69	3,450	3,899
QTR	15,525	3,814	159.85	3,465	4,091
SLP	51,320	2,400	65.13	2,301	2,556
SIN	38,462	3,300	49.69	3,218	3,413
SON	53,960	4,000	42.20	3,935	4,100
TAB	35,444	2,143	71.07	2,028	2,308
TAM	52,717	3,786	62.37	3,661	3,906
TLA	27,518	2,513	38.39	2,378	2,528
VER	166,179	2,138	29.97	2,020	2,138
YUC	49,913	2,228	51.55	2,143	2,345
ZAC	37,075	2,142	40.92	2,058	2,219
NACIONAL	**2,050,294**	**3,085**	**9.24**	**3,068**	**3,105**

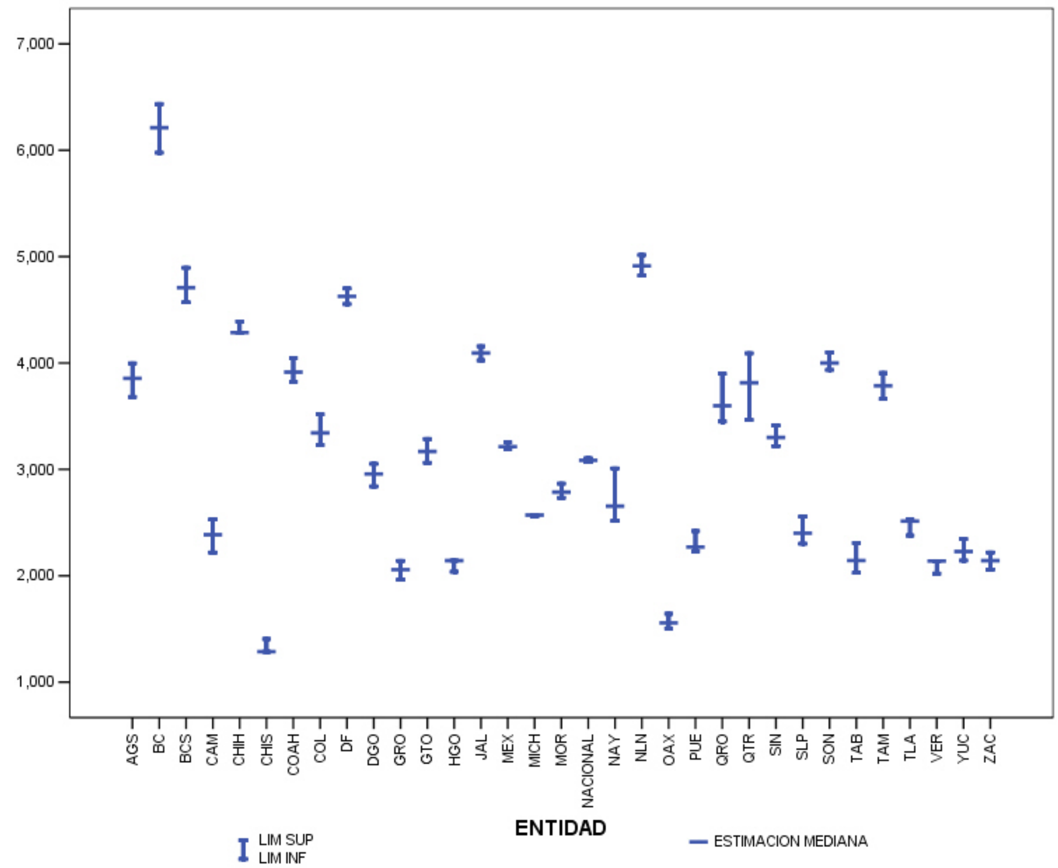

Figura 1.1: Estimación de la Mediana por Entidad Federativa.

Capítulo 2

Estimación del Coeficiente de Gini

2.1. La Medición de la desigualdad

Un índice de desigualdad es una medida que resume la manera como se distribuye una variable entre un conjunto de individuos. En el caso particular de la desigualdad económica, la medición se asocia al ingreso (o al gasto) de las familias o personas. Así, si $y_1, y_2, \ldots, y_n$ representan los ingresos de un grupo de n individuos, el indicador de desigualdad se construye como función de las observaciones.

Una primera clasificación de los indicadores de desigualdad, dada por Medina (2001), los agrupa como sigue: **medidas positivas**, que son aquellas que no hacen referencia explícita a ningún concepto de bienestar social, y medidas **normativas**, que sí están basadas en una función de bienestar. Al primer grupo pertenecen los índices estadísticos que tradicionalmente se utilizan para analizar la dispersión de una distribución de frecuencias, en tanto que hay diversas medidas normativas que se han propuesto para el estudio de la concentración del ingreso y la salud.

Una de las medidas probablemente más utilizadas es el denominado **Coeficiente de Gini**. Este indicador, que se clasifica entre las medidas estadísticas para el análisis de la distribución del ingreso, no utiliza como parámetro de referencia el ingreso medio de la distribución -a diferencia de la desviación media, la varianza y el coeficiente de variación-, su construcción puede derivarse a partir de la curva de Lorenz.

2.2. Diferencia Media.

La diferencia media (no confundir con desviación media) se define en Kendall y Stuart (1976) como

$$\begin{aligned} \Delta &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |x-y|\, dF(x)\, dF(y) \qquad (2.1) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |x-y|\, f(x)\, f(y)\, dxdy. \end{aligned}$$

En el caso discreto se tienen dos fórmulas. La primera

$$\Delta = \frac{1}{N(N-1)} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} |x_j - x_k|\, f(x_j)\, f(x_k), \qquad j \neq k, \qquad (2.2)$$

que es la diferencia media sin repetición, o

$$\Delta = \frac{1}{N^2} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} |x_j - x_k|\, f(x_j)\, f(x_k), \qquad (2.3)$$

la diferencia media con repetición. La diferencia de estas dos expresiones está sólo en el divisor, el cual es insignificante si N es grande.

La diferencia media fue propuesta por Gini (1912), pero discutida por F. R. Helmert en los 1870s. Tiene cierta atracción teórica, siendo dependiente de la dispersión de la variación de valores entre ellos y no de las desviaciones de algún valor central. Además, la integral que la define (2.1) puede converger aún cuando la varianza no converge. Debido a que es esencialmente una medida de dispersión lineal en lugar de una cuadrática, converge siempre que la media exista. Kendall y Stuart señalan que es, sin embargo, más dificil de calcular que la desviación estándar, y la presencia de valores absolutos en las ecuaciones que la definen indica, como en la desviación media, la presencia de dificultades en la teoría muestral. Quizá se piense que este inconveniente puede ser superado mediante la definición de un coeficiente

$$E^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-y)^2\, dF(x)\, dF(y).$$

Esto, sin embargo no es más que dos veces la varianza

$$\begin{aligned} E^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left\{x^2 - 2xy + y^2\right\} dF(x)\, dF(y) \\ &= 2\mu_2' - 2\mu_1'^2 = 2\mu_2. \end{aligned}$$

donde $\mu_2' = \int_{-\infty}^{\infty} (x-a)^2 \, dF$, $\mu_1' = \int_{-\infty}^{\infty} (x-a) \, dF$ y $\mu_2 = \int_{-\infty}^{\infty} (x-\mu_1')^2 \, dF$ con a un punto arbitrario.

Esto muestra que la varianza puede ser definida como la mitad de la media al cuadrado de todas las posibles diferencias variadas, esto es, sin referencia a desviaciones de un valor central, la media.

2.3. Coeficiente de Concentración de Gini

El coeficiente de concentración de Gini está definido de la siguiente manera, Kendall y Stuart (1976),

$$R = \frac{\Delta}{2\mu} \tag{2.4}$$

y se genera de manera natural de la siguiente aproximación.

Escribiendo, como es usual,

$$F(x) = \int_{-\infty}^{x} f(u) \, du \tag{2.5}$$

función de densidad acumulativa, defínase

$$\Phi(x) = \frac{1}{\mu} \int_{-\infty}^{x} u f(u) \, du, \tag{2.6}$$

donde

$$\mu = \int_{-\infty}^{\infty} y dF(y) \neq 0. \tag{2.7}$$

$\Phi(x)$ existe sólo si μ existe. Como $F(x)$ varía de 0 a 1, $\Phi(x)$ varía también de 0 a 1. $\Phi(x)$ puede ser llamado el primer momento *incompleto,* y como función de x es conocida como la *distribución de primer momento* de $F(x)$.

Las ecuaciones (2.5) y (2.6) definen una relación entre las variables F y Φ en términos de ecuaciones paramétricas en x. La curva cuya ordenada y abscisa son Φ y F, respectivamente, es conocida como la curva de concentración o la curva de Lorenz (ver figura 2.1).

La curva de Lorenz es una medida que fue propuesta en 1905 con el propósito de ilustrar la desigualdad en la distribución de la salud, Medina (2001), y, desde su aparición, su uso se ha popularizado entre los estudiosos de la desigualdad económica.

En términos simples, la curva de Lorenz representa el porcentaje acumulado de ingreso recibido por un determinado grupo de población ordenado en forma ascendente de acuerdo a la cuantía de su ingreso.

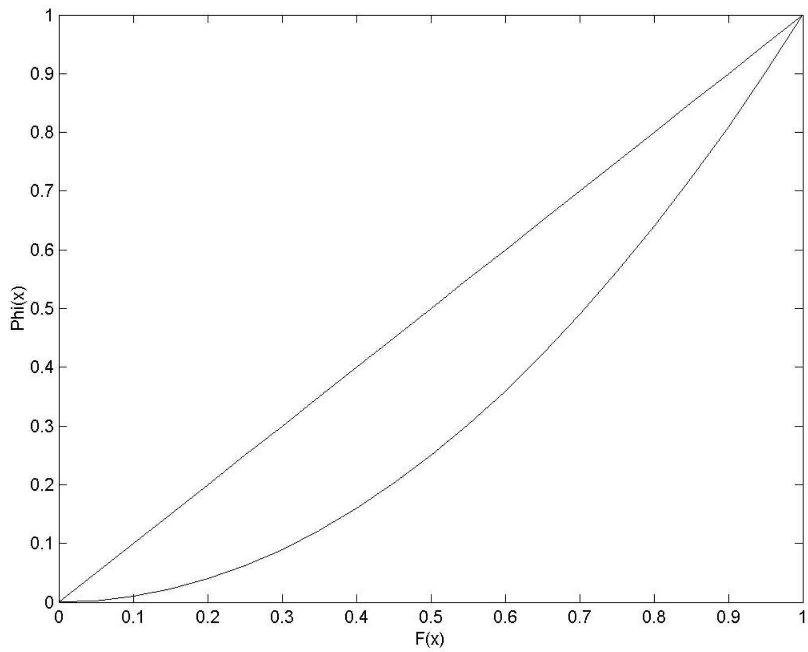

Figura 2.1: Curva de Concentración de Lorenz.

La curva de concentración de Lorenz debe ser convexa al eje F, por lo que se tiene

$$\mu \frac{d\Phi}{dF} = \mu \frac{\frac{d\Phi}{dx}}{\frac{dF}{dx}} = \frac{xf\left(x\right)}{f\left(x\right)} = x, \tag{2.8}$$

que es positiva, ya que el origen es tomado a la izquierda del inicio de la distribución. También

$$\mu \frac{d^2\Phi}{dF^2} = \frac{dx}{dF} = \frac{1}{f\left(x\right)} > 0.$$

De esta manera, la tangente a la curva hace un ángulo positivo con el eje F, y el ángulo se incrementa conforme F crece; en otras palabras, la curva es convexa al eje F.

La distancia entre la línea $\Phi = F$ y la curva de concentración es $F - \Phi\left(F\right)$. El punto crítico se tiene cuando $\frac{d}{dF}\left\{F - \Phi\right\} = 0$, es decir $d\Phi/dF = 1$. De (2.8), se tiene que este punto es en $x = \mu$; en la media, donde la curva de concentración es paralela a $\Phi = F$ y se da la mayor distancia entre ellas. Esta distancia máxima es

$$F\left(\mu\right) - \Phi\left(\mu\right) = -\frac{1}{\mu}\int_{-\infty}^{\mu}\left(u-\mu\right)dF\left(u\right) = \frac{1}{2}\frac{\delta}{\mu}$$

donde $\delta = \int_{-\infty}^{\infty} |x-\mu|\, dF$ es la desviación media.

El área entre la curva de concentración y la recta $\Phi = F$ es llamada el área de concentración. A continuación se muestra que esta área es igual a un medio el coeficiente de concentración.

De la figura 2.1 se tiene que ($A =$ área de concentración)

$$2A = \int_0^1 F d\Phi - \int_0^1 \Phi dF \tag{2.9}$$

y de esta manera, de (2.8) se tiene

$$d\Phi(x) = \frac{1}{\mu} x dF(x)$$

por lo que (2.9) se convierte en

$$\begin{aligned} 2\mu A &= \int_0^\infty F(x)\, x dF(x) - \mu \int_0^\infty \Phi(x)\, dF(x) \qquad (2.10) \\ &= \int_0^\infty x \int_0^x dF(y)\, dF(x) - \int_0^\infty \int_0^x y dF(y)\, dF(x) \\ &= \int_0^\infty \int_0^x (x-y)\, dF(y)\, dF(x)\,. \end{aligned}$$

Ahora $\int_0^\infty \int_0^\infty (x-y)\, dF(y)\, dF(x) = 0$, y de aquí

$$\begin{aligned} 2\mu A &= \frac{1}{2}\left[\int_0^\infty \int_0^x (x-y)\, dF(y)\, dF(x) + \int_0^\infty \int_x^\infty (y-x)\, dF(y)\, dF(x)\right] \\ &= \frac{1}{2}\int_0^\infty \int_0^\infty |x-y|\, dF(x)\, dF(y) = \frac{1}{2}\Delta. \end{aligned}$$

Entonces, el área de concentración es igual a $\frac{1}{4}\Delta/\mu$, esto es, un medio del coeficiente de concentración R (2.4). De la figura 2.1, el área está entre 0 y $\frac{1}{2}$, así que $0 \leq R \leq 1$.

Considerando la ecuación (2.10), se puede ver que una definición alternativa a la diferencia media es

$$\Delta = 2\int_0^\infty \{xF(x) - \mu\Phi(x)\}\, dF(x)\,. \tag{2.11}$$

La interpretación que se obtiene de la curva de Lorenz es que cuando está más cercana a la diagonal de 45°, en el cuadrado unitario, la repartición de la riqueza es más igualitaria dentro de la población de estudio, caso contrario ocurre cuando esta curva se aproxima al eje de las abscisas (a cero). No se ilustra la obtención de la curva de Lorenz a datos muestrales debido a que no es el objetivo principal de este trabajo.

El coeficiente de Gini puede ser usado para medir la dispersión de una distribución de ingreso, o consumo, o riqueza, o una distribución de cualquier otro tipo. Usualmente, el índice de Gini se usa mayormente en la distribución del ingreso. Una distribución de ingresos quizá se tenga de diferentes medios: ingresos de la cabeza de familia o ingresos individuales. El concepto de distribución del ingreso puede variar de los considerados antes de impuestos y otras transferencias fiscales a los que son considerados después de estos movimientos. En el presente trabajo no se consideran tales detalles, se está enfocado en la forma operativa del cálculo de este índice.

El índice de Gini tiene diferentes formulaciones e interesantes interpretaciones. Se puede expresar como un cociente de dos regiones definidas por una línea de 45° y la curva de Lorenz en un cuadrado unitario, o una función de la diferencia media de Gini, o una covarianza entre ingresos y sus rangos, o una forma matricial de un tipo especial.

Una expresión que se propone al cálculo del índice de Gini para una población finita es la dada por Nygård y Sandström (1985) y que tiene la siguiente forma

$$R_N = \frac{2}{N^2 \bar{y}_N} \sum_{i=1}^{N} i y_{(i)} - 1 - \frac{1}{N}. \tag{2.12}$$

donde $\bar{y}_N = \frac{1}{N} \sum_{k=1}^{N} y_k$; $y_{(i)}$ denota la $i - ésima$ estadística de orden y N es el total de la población. En trabajos posteriores al de Nygård y Sandström, Sandström et al. (1988) presenta una estimación a este coeficiente y propone una estimación de su varianza. Más adelante se hará una revisión detallada de tal trabajo.

2.4. La Técnica Jackknife

La técnica jackknife se originó fuera del campo de muestreo de encuestas. La primer idea, desarrollada por Quenouille (1949, 1956), fue usar jackknife para reducir el sesgo de un estimador en un contexto de población infinita. Tukey (1958) sugirió subsecuentemente que la técnica debería también ser usada para producir estimaciones de varianza. Para poblaciones finitas, la técnica jackknife fue considerada primero por Durbin (1959). A continuación se da una revisión de la técnica jackknife como es comúnmente usada para estimar una varianza en un muestreo de encuestas. Para más detalle referirse a Särndal et al. (1992).

Sea s una muestra (la muestra completa) de n elementos obtenidos de un diseño muestral de elementos arbitrarios no estratificados (muestreo estratificado se considerará posteriormente). Sea θ el parámetro poblacional a ser estimado por $\hat{\theta}$, un estimador basado en datos de la muestra completa s. El propósito es estimar $V\left(\hat{\theta}\right)$.

En Särndal et al. (1992) se describe a detalle el procedimiento de esta técnica, mencionada a detalle a continuación. La técnica jackknife inicia con la partición de la muestra s en A grupos aleatorios dependientes de igual tamaño $m\,(=n/A)$. Se asume que, para cualquier s dado, cada grupo es un muestreo aleatorio simple (m.a.s.) de s, aún si s no es un m.a.s. Después, para cada grupo $(a=1,\ldots,A)$, se calcula $\hat{\theta}_{(a)}$, un estimador de θ de la misma forma funcional que $\hat{\theta}$, pero basado sólo en los datos que permanecen después de omitir el $a-\acute{e}simo$ grupo. Para $a=1,\ldots,A$, se define

$$\hat{\theta}_a = A\hat{\theta} - (A-1)\,\hat{\theta}_{(a)} \tag{2.13}$$

algunas veces llamado el $a-\acute{e}simo$ pseudovalor. El estimador jackknife de θ (alternativa al estimador $\hat{\theta}$) es

$$\hat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_a \tag{2.14}$$

y el estimador de varianza jackknife está definido como

$$\widehat{V}_{JK1} = \frac{1}{A\,(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}_{JK}\right)^2. \tag{2.15}$$

En la práctica, $\widehat{V}_{JK1}$ es usado como un estimador de $V\left(\hat{\theta}\right)$ así como de $V\left(\hat{\theta}_{JK}\right)$. Un estimador alternativo de $\widehat{V}_{JK1}$ es

$$\widehat{V}_{JK2} = \frac{1}{A\,(A-1)}\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}\right)^2. \tag{2.16}$$

Nótese que $\widehat{V}_{JK2} \geq \widehat{V}_{JK1}$, debido a que

$$\sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}\right)^2 = \sum_{a=1}^{A}\left(\hat{\theta}_a - \hat{\theta}_{JK}\right)^2 + A\left(\hat{\theta}_{JK} - \hat{\theta}\right)^2.$$

En general, las $\hat{\theta}_a$ están correlacionadas, así que el insesgamiento no se obtiene. No hay resultados exactos (tamaño de muestra finito) para las propiedades de los estimadores de varianza jackknife cuando θ es más complejo que un estimador π[1]. En el caso simple, cuando θ es un total poblacional y $\hat{\theta}$ el correspondiente estimador π, la técnica jackknife trabaja bien. Este resultado, junto con la evidencia empírica, parece ser la principal justificación para el uso de la técnica jackknife en situaciones más complejas.

La Técnica Jackknife en Muestreos Estratificados

Cuando se aplica la técnica jackknife a un muestreo estratificado, usualmente se usan otros estimadores de varianza diferentes a (2.15) y (2.16). De acuerdo a Wolter (1985, p. 175), uno debería ser "especialmente cuidadoso de no aplicar los estimadores jackknife clásicos a problemas de muestreo estratificado."

Asuma que la muestra de H estratos es particionada en A_h grupos ($h = 1, \ldots, H$), de un total de $A = \sum_{h=1}^{H} A_h$ grupos. Como antes, sea $\hat{\theta}$ el estimador de θ de la muestra completa. Sea $\hat{\theta}_{(ha)}$ el estimador de θ basado en la muestra restante del estrato h después de omitir el $a-ésimo$ grupo. Un estimador que se sugiere en Särndal et al. (1992) para $V\left(\hat{\theta}\right)$ es

$$\widehat{V}_{JK3} = \sum_{h=1}^{H} \frac{A_h - 1}{A_h} \sum_{a=1}^{A_h} \left(\hat{\theta}_{(ha)} - \hat{\theta}\right)^2 \tag{2.17}$$

el cual sería insesgado si la selección de la muestra fuera con reemplazo y si $\hat{\theta}$ fuera el estimador π del total de la población $\theta = t = \sum_U y_k$.

En particular, suponga que un m.a.s. es usado en cada estrato. Sea

$$\hat{\theta} = \sum_{h=1}^{H} N_h \bar{y}_{s_h}$$

y

$$\hat{\theta}_{(ha)} = N_1 \bar{y}_{s_1} + \cdots + N_{h-1} \bar{y}_{s_{h-1}} + N_h \bar{y}_{s_h - s_{ha}} + N_{h+1} \bar{y}_{s_{h+1}} + \cdots + N_H \bar{y}_{s_H}.$$

Entonces se obtiene

$$\widehat{V}_{JK3} = \sum_{h=1}^{H} \frac{N_h^2}{A_h \left(A_h - 1\right)} \sum_{a=1}^{A_h} \left(\bar{y}_{s_{ha}} - \bar{y}_{s_h}\right)^2.$$

[1]El estimador π, de un total, es el que considera los factores de expansión de un muestreo general, $\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$.

Bajo el suspuesto de que $A_h = n_h$, se obtiene simplemente

$$\widehat{V}_{JK3} = \sum_{h=1}^{H} \frac{N_h^2}{n_h} S_{ys_h}^2$$

el cual es el estimador de varianza tradicional, omitiendo el factor de corrección de la población por finitud.

La Técnica Jackknife en Muestreos Multietápicos

En muestreo multietápico, la técnica jackknife es aplicada usualmente al nivel de las Unidades Primarias de Muestreo (UPM). Supóngase un muestreo de primera etapa s_I que contiene n_I UPM's obtenidas del conjunto U_I compuesto por N_I UPM's. Sea s_I particionada en A grupos aleatorios de las UPM's, con m UPM's en cada grupo. Sea $\hat{\theta}$ el estimador de θ de la muestra completa, y sea $\hat{\theta}_{(a)}$ que denota el estimador de θ basado en los datos restantes después de eliminar el $a-\acute{e}simo$ grupo de UPM's. Usando las ecuaciones (2.13) a (2.16) en este nuevo contexto, se obtiene

$$\begin{aligned} \hat{\theta}_a &= A\hat{\theta} - (A-1)\,\hat{\theta}_{(a)}; \qquad a = 1, \ldots, A \\ \hat{\theta}_{JK} &= \frac{1}{A} \sum_{a=1}^{A} \hat{\theta}_a \end{aligned}$$

y los estimadores de la varianza son

$$\begin{aligned} \widehat{V}_{JK1} &= \frac{1}{A\,(A-1)} \sum_{a=1}^{A} \left(\hat{\theta}_a - \hat{\theta}_{JK}\right)^2 \\ \widehat{V}_{JK2} &= \frac{1}{A\,(A-1)} \sum_{a=1}^{A} \left(\hat{\theta}_a - \hat{\theta}\right)^2. \end{aligned}$$

los cuales sirven como propósito dual de estimar $V\left(\hat{\theta}\right)$ así como de $V\left(\hat{\theta}_{JK}\right)$.

Para jackknife se necesita un número de grupos A, fijo, adecuados. Para una mejor exactitud en el resultado del estimador de varianza se desearía tener tantos grupos como sea posible, lo cual significa $A = n$, esto es $m = 1$. Por otro lado, por razones computacionales, se debería preferir los menos grupos como sea posible, el caso más extremo es $A = 2$, y $m = n/2$. En la práctica, la elección más frecuente parece ser $A = n$, aunque queda a consideración del analista el número de grupos que desee hacer de su muestra, dependiendo de las características del diseño de la misma, para así aplicar la técnica jackknife y tener resultados óptimos.

2.5. Estimador del Indice de Gini que ignora Diseño Muestral

En este trabajo, se presentan dos métodos para la estimación del índice de Gini y de su varianza. Uno de ellos es considerando diseño muestral simple y el otro bajo un diseño muestral complejo, que también es usado para obtener una estimación suponiendo un m.a.s.

El primer método mencionado se atribuye a Ogwang (2000), el cual está diseñado para muestreos aleatorios simples ya que no considera ningún diseño muestral complejo en la estimación del índice de Gini ni de su varianza. Para la estimación del índice se hace uso, básicamente, de los Mínimos Cuadrados Ordinarios (MCO) y para el caso del error estándar se propone un algoritmo sencillo usando la técnica jackknife, vista a detalle en la sección §2.4. A continuación se presenta una breve descripción de este método.

El índice de Gini definido por Ogwang tiene la forma:

$$R_N = \frac{1}{\mu\left(x\right)} \int_a^b F\left(x\right) \left[1 - F\left(x\right)\right] dx \tag{2.18}$$

donde a y b denotan el ingreso más pequeño y el más grande, respectivamente, $\mu\left(x\right)$ es la media del ingreso, y $F\left(x\right)$ es la distribución acumulativa de los ingresos. Ogwang mostró que si los ingresos muestrales se ordenan en forma ascendente, con $x_{(i)}$ que denota el $i - \acute{e}simo$ componente más pequeño, entonces

$$\hat{R}_N = \left(\frac{n^2 - 1}{6n}\right) \frac{\hat{\beta}}{\bar{x}} \tag{2.19}$$

donde $\hat{\beta}$ es el estimador de MCO de β en el modelo de regresión

$$x_{(i)} = \alpha + \beta i + u_{(i)} \qquad i = 1, \ldots, n \tag{2.20}$$

en el cual se asume que $E\left(u_{(i)}\right) = 0$ y $Var\left(u_{(i)}\right) = \sigma^2$.

Una forma alternativa más conveniente de la ecuación (2.19) es:

$$\hat{R}_N = \frac{2}{n}\hat{\alpha} - 1 - \frac{1}{n} \tag{2.21}$$

donde

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} i x_{(i)}}{\sum_{i=1}^{n} x_i}$$

es el estimador de mínimos cuadrados ponderados (MCP) de α en el modelo

$$i = \alpha + u_i \qquad i = 1, \ldots, n \tag{2.22}$$

suponiendo ahora que $E\left(u_i^2\right)=\frac{\sigma^2}{x_i}$.

La ecuación (2.21), la cuál se usa en la derivación de un algoritmo simple para el cálculo del error estándar jackknife del índice de Gini (véase sección §2.4), muestra que el índice de Gini es lineal en $\hat{\alpha}$.

A pesar de que la mayoría de los errores estándar jackknife de desigualdades del ingreso son relativamente fáciles de calcular, el índice de Gini no es el caso, sobretodo cuando se tienen muchas observaciones. La varianza jackknife del índice de Gini está dada por la siguiente fórmula

$$\widehat{V}\left(\hat{R}_N\right)=\left(\frac{n-1}{n}\right)\sum_{k=1}^{n}\left(\hat{R}_N\left(n,k\right)-\hat{\bar{R}}_N\left(n\right)\right)^2 \tag{2.23}$$

donde $\hat{R}_N\left(n,k\right)$ es el pseudo-estimador del índice de Gini de las $(n-1)$ observaciones al no considerar la $k-\acute{e}sima$; $\hat{\bar{R}}_N\left(n\right)=\sum_{k=1}^{n}\hat{R}_N\left(n,k\right)/n$ es la media de todas las $\hat{R}_N\left(n,k\right)'s$, $k=1,\ldots,n$.

Para derivar un algoritmo de la estimación del error estándar jackknife del índice de Gini, se denota $\hat{R}_N\left(n,0\right)$ como el estimador del índice de Gini cuando se consideran las n observaciones. Además, se considera que los ingresos son ordenados en forma ascendente. En Ogwang (2000) se presenta la siguiente igualdad, que es equivalente a la expresión (2.21)

$$\begin{aligned}\hat{R}_N\left(n,k\right) &= \hat{R}_N\left(n,0\right)+\left(\frac{2}{n}\right)\left[x_k\hat{\alpha}\left(n,0\right)+\frac{\sum\limits_{i=1}^{n}ix_{(i)}}{n-1}\right]\left[\frac{1}{\sum\limits_{i=1}^{n}x_i-x_k}\right]\\ &\quad-\left(\frac{2}{n-1}\right)\left[\sum_{i=1}^{n}x_i-\sum_{i=1}^{k}x_{(i)}+kx_k\right]\left[\frac{1}{\sum\limits_{i=1}^{n}x_i-x_k}\right]\\ &\quad-\frac{1}{n\left(n-1\right)}\end{aligned} \tag{2.24}$$

donde $\hat{\alpha}\left(n,0\right)=\sum_{i=1}^{n}ix_{(i)}/\sum_{i=1}^{n}x_i$ es el estimador de MCP de α en la ecuación (2.22) cuando todas las observaciones son consideradas.

Esta última igualdad facilita en forma considerable la estimación del error estándar jackknife para el índice de Gini propuesto por Owgang.

2.6. Estimador del Indice de Gini que considera Diseño Muestral

Un estimador del índice de Gini (2.4) es el propuesto por Sandström et al. (1988), pág. 113-119 [21], el cual considera los factores de expansión del diseño muestral considerado. La expresión analítica que proponen tiene la siguiente forma:

$$\hat{R}_N = \frac{\sum\limits_{k \in s} \left(2\hat{P}_N\left(y_k\right) + \pi_k^{-1}\right) y_k \pi_k^{-1}}{\hat{N}^2 \widehat{\bar{y}}_N} - 1 \tag{2.25}$$

donde

$$\begin{aligned} \hat{P}_N\left(y\right) &= \sum_{k \in s} I\left\{y_k \leq y\right\} \pi_k^{-1} \\ \widehat{\bar{y}}_N &= \frac{1}{\hat{N}} \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{\sum_{k \in s} \frac{y_k}{\pi_k}}{\sum_{k \in s} \frac{1}{\pi_k}} = \frac{\hat{t}_{y\pi}}{\hat{N}} \\ I\left\{y_k \leq y\right\} &= \begin{cases} 1; & y_k \leq y \\ 0; & y_k > y \end{cases} \end{aligned}$$

Obsérvese que para el caso de un m.a.s., donde $\pi_k = \frac{n}{N}, k = 1, \ldots, N$, la fórmula (2.25) se reduce a lo siguiente:

$$\hat{R}_{Nmas} = \frac{2 \sum\limits_{k \in s} y_k \hat{P}_{Nmas}\left(y_k\right)}{\bar{y}_s} - 1 + \frac{1}{n} \tag{2.26}$$

donde

$$\begin{aligned} \hat{P}_{Nmas}\left(y_k\right) &= \sum_{k \in s} I\left\{y_k \leq y\right\} \\ \bar{y}_s &= \frac{1}{n} \sum_{k \in s} y_k. \end{aligned}$$

2.6.1. Estimador de Varianza del Coeficiente de Gini considerando Diseño Muestral

Sea σ^2 la varianza muestral de $\hat{R}_N$. Como $\hat{R}_N$ es un estadístico no lineal, su varianza no se puede expresar en una forma analítica simple, tampoco es posible obtener su estimación por los métodos tradicionales de estimación de varianza insesgada conocidos en la teoría de muestreo de encuestas. En este caso, se procederá a obtener la estimación de la varianza mediante técnicas de aproximación.

Estimador propuesto por Sandström et al. (1988). Una expresión analítica para el estimador de la varianza del índice de Gini dado por la ecuación (2.25) es un estimador consistente del valor asintótico de $\mathbb{E}\left[\left(\hat{R}_N - R_N\right)^2\right]$, donde $\mathbb{E}$ denota el valor esperado con respecto a la distribución conjunta de variables aleatorias i.i.d. $Y_1, \ldots, Y_N$, condicionado a una muestra s dada, y que está dada en forma explícita por la siguiente expresión:

$$\widehat{V}\left(\hat{R}_N\right) = \left[\frac{1 - \hat{f} + \hat{v}^2}{n\widehat{\overline{y}}_N^2}\right]\hat{\sigma}^2 \tag{2.27}$$

donde

$$\begin{aligned}
\hat{f} &= \frac{n}{\hat{N}}; \qquad \hat{N} = \sum_{i=1}^{n}\frac{1}{\pi_i} \\
\hat{v}^2 &= \left(\frac{\hat{f}}{\hat{N}}\right)\sum_{i=1}^{n}\left(\frac{1}{\pi_i} - \frac{1}{\hat{f}}\right)^2 \\
\hat{\sigma}^2 &= \left(\left(1 + \hat{R}_N\right)\Delta_0 - 2\Delta_1\right)\left(\left(1 + \hat{R}_N\right)\Delta_1 - 2\Delta_2\right) \\
&\quad - \left(\left(1 + \hat{R}_N\right)\Delta_1 - 2\Delta_2\right)^2 \\
&\quad - \frac{1}{2}\left\{4S_2 - 4\left(1 + \hat{R}_N\right)S_1 + \left(1 + \hat{R}_N\right)^2 S_3\right\}
\end{aligned}$$

donde

$$\begin{aligned}
\Delta_0 &= y_{(n)} - y_{(1)} \\
\Delta_1 &= y_{(n)} - \widehat{\overline{y}}_N \\
\Delta_2 &= y_{(n)} - \widehat{\overline{y}}_N\left(1 + \hat{R}_N\right) \\
y_{(i)},\ i &= 1, ..., n \text{ estadísticas de orden.}
\end{aligned}$$

Además, sean

$$\begin{aligned}
\Delta_{0i} &= y_{(i)} - y_{(1)}, \qquad \text{i.e., } \Delta_{0n} = \Delta_0, \\
\Delta_{1i} &= F_i y_{(i)} - \frac{1}{\hat{N}}\sum_{\left\{y_k \le y_{(i)}\right\}}\frac{y_k}{\pi_k}, \qquad \text{i.e., } \Delta_{1n} = \Delta_1, \\
\Delta_{2i} &= F_i^2 y_{(i)} - \sum_{\left\{y_k \le y_{(i)}\right\}}\left(F_k^2 - F_{k-1}^2\right)y_k, \qquad \text{i.e., } \Delta_{2i} = \Delta_2,
\end{aligned}$$

donde

$$F_i = \frac{1}{\hat{N}}\sum_{j=1}^{n}\frac{I\left\{y_j \le y_i\right\}}{\pi_j}$$

Ahora, S_1, S_2, y S_3 son tales que los índices de las sumas toman valores $i = 1, ..., n$

$$\begin{aligned}
S_1 &= \Delta_1^2 - \Delta_2\Delta_0 + 2y_{(n)}\left(\Delta_0 - \Delta_1\right) \\
&\quad -2\sum F_i^2\left(\Delta_{0i} - \Delta_{0i-1}\right)y_i \\
&\quad +\frac{2}{\hat{N}^2}\sum \Delta_{0i-1}\frac{y_i}{\pi_i^2} - \frac{4}{\hat{N}}\sum F_i\Delta_{0i-1}\frac{y_i}{\pi_i} \\
&\quad +2\sum F_i\left(\Delta_{1i} - \Delta_{1i-1}\right)y_i + \frac{2}{\hat{N}}\sum \Delta_{1i-1}\frac{y_i}{\pi_i},
\end{aligned}$$

$$\begin{aligned}
S_2 &= 2y_{(n)}\left(\Delta_1 - \Delta_2\right) + 2\sum F_i\left(\Delta_{2i} - \Delta_{2i-1}\right)y_i \\
&\quad -2\sum F_i^2\left(\Delta_{1i} - \Delta_{1i-1}\right)y_i + \frac{2}{\hat{N}^2}\sum \Delta_{1i-1}\frac{y_i}{\pi_i^2} \\
&\quad +\frac{2}{\hat{N}}\sum \Delta_{2i-1}\frac{y_i}{\pi_i} - \frac{4}{\hat{N}}\sum F_i\Delta_{1i-1}\frac{y_i}{\pi_i}.
\end{aligned}$$

$$\begin{aligned}
S_3 &= 2y_{(n)}\left(\Delta_0 - \Delta_1\right) - 2\sum F_i\left(\Delta_{0i} - \Delta_{0i-1}\right)y_i \\
&\quad -\frac{2}{\hat{N}}\sum \Delta_{0i-1}\frac{y_i}{\pi_i} + 2\sum\left(\Delta_{1i} - \Delta_{1i-1}\right)y_i.
\end{aligned}$$

Este estimador de la varianza del Coeficiente de Gini fue comparado con otros dos estimadores obtenidos mediante: el primero basado en la técnica jackknife, el cual se menciona adelante con más detalle, y el segundo mediante aproximación de series de Taylor de primer orden. En Sandström et al. (1988) se menciona que el estimador que propusieron, el dado en la expresión (2.27), es el que dio mejores resultados de estimación al ejemplo de estudio que consideraron.

Obsérvese que para el caso de un m.a.s., donde $\pi_k = \frac{n}{N}, k = 1, \ldots, N$, la expresión (2.27) se reduce a lo siguiente:

$$\widehat{V}\left(\hat{R}_{Nmas}\right) = \left(\frac{1-f}{n\bar{y}_s^2}\right)\hat{\sigma}^2$$

donde

$$f = \frac{n}{N}; \qquad \bar{y}_s = \sum_{i=1}^{n}\frac{y_k}{n}$$

$\hat{\sigma}^2$ definida como arriba, pero considerando $\hat{R}_{Nmas}$ en vez de $\hat{R}_N$ y las modificaciones en sus factores:

$$\begin{aligned}
\Delta_0 &= y_{(n)} - y_{(1)} \\
\Delta_1 &= y_{(n)} - \bar{y}_s \\
\Delta_2 &= y_{(n)} - \bar{y}_s\left(1 + \hat{R}_{Nmas}\right) \\
y_{(i)},\ i &= 1, ..., n \text{ estadísticas de orden.}
\end{aligned}$$

Además de los cambios en los siguientes factores

$$
\begin{aligned}
\Delta_{0i} &= y_{(i)} - y_{(1)}, \qquad \text{i.e., } \Delta_{0n} = \Delta_0, \\
\Delta_{1i} &= F_i y_{(i)} - \frac{1}{n} \sum_{\left\{y_k \leq y_{(i)}\right\}} y_k, \qquad \text{i.e., } \Delta_{1n} = \Delta_1, \\
\Delta_{2i} &= F_i^2 y_{(i)} - \sum_{\left\{y_k \leq y_{(i)}\right\}} \left(F_k^2 - F_{k-1}^2\right) y_k, \qquad \text{i.e., } \Delta_{2i} = \Delta_2,
\end{aligned}
$$

donde

$$
F_i = \frac{1}{n} \sum_{j=1}^{n} I\left\{y_j \leq y_i\right\}
$$

Ahora, $S_1, S_2,$ y S_3 se alteran de la siguiente manera, los índices de las sumas toman valores $i = 1, ..., n$

$$
\begin{aligned}
S_1 &= \Delta_1^2 - \Delta_2 \Delta_0 + 2 y_{(n)} \left(\Delta_0 - \Delta_1\right) \\
&\quad - 2 \sum F_i^2 \left(\Delta_{0i} - \Delta_{0i-1}\right) y_i \\
&\quad + \frac{2}{n^2} \sum \Delta_{0i-1} y_i - \frac{4}{n} \sum F_i \Delta_{0i-1} y_i \\
&\quad + 2 \sum F_i \left(\Delta_{1i} - \Delta_{1i-1}\right) y_i + \frac{2}{n} \sum \Delta_{1i-1} y_i,
\end{aligned}
$$

$$
\begin{aligned}
S_2 &= 2 y_{(n)} \left(\Delta_1 - \Delta_2\right) + 2 \sum F_i \left(\Delta_{2i} - \Delta_{2i-1}\right) y_i \\
&\quad - 2 \sum F_i^2 \left(\Delta_{1i} - \Delta_{1i-1}\right) y_i + \frac{2}{n^2} \sum \Delta_{1i-1} y_i \\
&\quad + \frac{2}{n} \sum \Delta_{2i-1} y_i - \frac{4}{n} \sum F_i \Delta_{1i-1} y_i.
\end{aligned}
$$

$$
\begin{aligned}
S_3 &= 2 y_{(n)} \left(\Delta_0 - \Delta_1\right) - 2 \sum F_i \left(\Delta_{0i} - \Delta_{0i-1}\right) y_i \\
&\quad - \frac{2}{n} \sum \Delta_{0i-1} y_i + 2 \sum \left(\Delta_{1i} - \Delta_{1i-1}\right) y_i.
\end{aligned}
$$

2.6.2. Estimador de Varianza usando la Técnica Jackknife

El estimador siguiente, $\hat{\sigma}_{JK}^2$, está basado en la técnica jackknife, visto a detalle en la sección §2.4. Una observación no es considerada en la muestra a la vez. Cada vez que se borra, se calcula $\hat{R}_N^{(j)}$, análogo a $\hat{R}_N$, basado en las restantes $n-1$ observaciones y borrando la $j-\acute{e}sima$ observación, $j = 1, 2, \ldots, n$. La fórmula de la estimación de la varianza es similar a la dada en (2.23)

$$
\hat{\sigma}_{JK}^2 = \left(\frac{n-1}{n}\right) \sum_{j=1}^{n} \left(\hat{R}_N^{(j)} - \hat{R}_N^{(\cdot)}\right)^2, \tag{2.28}
$$

donde

$$\hat{R}_N^{(\cdot)} = \frac{1}{n}\sum_{j=1}^{n}\hat{R}_N^{(j)}.$$

Para la obtención de la expresión $\hat{R}_N^{(j)}$ se actualizaron en cada paso los factores de expansión π_k^{-1} a otros $\left(\pi_k^{(j)}\right)^{-1}$ de tal manera que se siga conservando la propiedad de que

$$\hat{N} = \sum_{k=1}^{n}\frac{1}{\pi_k}$$

por lo que las probabilidades de inclusión deben ser actualizadas de la siguiente manera, esto es considerando que se tiene un diseño muestral diferente del aleatorio simple,

$$\begin{aligned}\pi_k^{(j)} &= \pi_k\frac{\sum_{\substack{k=1\\k\neq j}}^{n}\pi_k^{-1}}{\sum_{k=1}^{n}\pi_k^{-1}}\\ &= \pi_k\frac{\hat{N}^{(j)}}{\hat{N}}\end{aligned}$$

donde

$$\hat{N}^{(j)} = \sum_{\substack{k=1\\k\neq j}}^{n}\frac{1}{\pi_k} = \hat{N} - \frac{1}{\pi_j}, \quad \text{y} \quad \hat{N} = \sum_{k=1}^{n}\frac{1}{\pi_k}$$

quedando los factores de expansión actualizados como

$$\frac{1}{\pi_k^{(j)}} = \frac{1}{\pi_k}\frac{\hat{N}}{\hat{N}^{(j)}}.$$

Una vez actualizados los factores de expansión, la estimación de todas las $\hat{R}_N^{(j)};\ j = 1,2,\ldots,n$ puede ser engorrosa, sobretodo cuando se tiene un tamaño de muestra n muy grande. Es posible demostrar que la siguiente igualdad se cumple

$$\begin{aligned}\hat{R}_N^{(j)} &= \frac{\sum_{k\in s^{(j)}}\left(2\hat{P}_N^{(j)}(y_k) + \frac{1}{\pi_k^{(j)}}\right)\frac{y_k}{\pi_k^{(j)}}}{\hat{N}^{(j)2}\widehat{\overline{y}}_N^{(j)}} - 1\\ &= \frac{\sum_{k\in s^{(j)}}\left[2\left(\hat{P}_N(y_k) - I\{y_k > y_j\}\frac{1}{\pi_j}\right) + \frac{1}{\pi_k}\right]\frac{y_k}{\pi_k}}{\hat{N}^{(j)}\left(\hat{N}\widehat{\overline{y}}_N - \frac{y_j}{\pi_j}\right)} - 1\\ &= \frac{\sum_{k\in s^{(j)}}\left[2\left(\hat{P}_N(y_k) - I\{y_k > y_j\}\frac{1}{\pi_j}\right) + \frac{1}{\pi_k}\right]\frac{y_k}{\pi_k}}{\left(\hat{N} - \frac{1}{\pi_j}\right)\left(\sum_{k\in s}\frac{y_k}{\pi_k} - \frac{y_j}{\pi_j}\right)} - 1,\end{aligned}$$

donde $s^{(j)}$ es la muestra sin considerar el $j-ésimo$ elemento y s es la muestra con todos los elementos.

En la última expresión, $\hat{R}_N^{(j)}$ sólo depende de valores calculados previamente en la obtención de $\hat{R}_N$ (considerando las n observaciones), es decir, no es necesaria la actualización de las probabilidades de inclusión en cada paso ($j = 1, \ldots, n$) y por consiguiente los factores de expansión en cada paso, siendo una expresión que facilita la estimación de la varianza jackknife en forma considerable.

2.7. Resultados Numéricos Obtenidos y Conclusiones

A continuación se presentan los resultados de los métodos mencionados anteriormente en la estimación del coeficiente de Gini y de su error estándar.

Para ilustrar el método de estimación del índice de Gini, y el cálculo de su intervalo de confianza, se utilizó la misma información que para la estimación de la Mediana, esto es, la variable de Ingreso Mensual Promedio en Hogares para información proveniente de la muestra del XII Censo de Población y Vivienda 2000. El esquema de muestreo que presenta el Censo es por conglomerados en una sola etapa, es decir, se seleccionaron áreas geográficas completas, ya fueran áreas geoestadísticas básicas (Ageb), manzanas o localidades rurales. La selección de espacios geográficos completos permite aplicar un solo cuestionario en las viviendas: el ampliado en las seleccionadas y el básico en el resto. De esta manera, el total de la población se obtiene de sumar la información proveniente de los cuestionarios básico y ampliado.

En la tabla 2.1 se muestran los resultados obtenidos mediante el método propuesto por Ogwang. En este método no se considera un diseño muestral específico para la estimación de los parámetros, ya que como se mencionó en la sección §2.5, la estimación del índice de Gini se hace mediante el uso de los Mínimos Cuadrados Ponderados mientras que la estimación de su varianza se hace utilizando la técnica jackknife.

En la tabla 2.2 los resultados obtenidos fueron generados a partir de los estimadores propuestos por Sandström et al. (ecuaciones 2.25 y 2.27), pero considerando un diseño muestral aleatorio simple, esto es $\pi_k = \frac{n}{N}$, $\forall k \in s$. Aquí cabe comentar que los resultados obtenidos mediante este método fueron muy similares a los presentados por el método de Ogwang (véase gráfica 2.2), tanto en la estimación puntual como en la estimación del error estándar, siendo ligeramente más grande el error estándar obtenido mediante el método de Ogwang para todos los estados. Esto se aprecia con mayor claridad en la tabla 2.5 donde en la columna **A/B** están los cocientes por entidad de la estimación del error estándar obtenido por el método de Ogwang, numerador, y el de Sandström et al., denominador, mostrando mayor magnitud las obtenidas por Ogwang.

A pesar de que los resultados obtenidos por el método propuesto por Ogwang y los obtenidos mediante las expresiones de Sandström, aplicadas a un muestreo aleatorio simple, son muy similares, no es posible determinar cual de los dos métodos da mejores estimaciones, ya que se desconoce el valor real del índice de Gini y su correspondiente varianza para así poder comparar los resultados obtenidos. Sólo se puede determinar que ambos métodos producen resultados de estimación puntual similares y de varianza muy parecidos, siendo de mayor magnitud las estimaciones de Ogwang.

En la tabla 2.3 se muestran los resultados de los estimadores propuestos por Sandström, considerando el diseño muestral complejo. Es decir, utilizando las probabilidades de inclusión, π_k, que fueron generadas a partir del diseño muestral. En este caso, la mayoría de los estimadores puntuales del coeficiente de Gini resultaron ser más pequeños que los obtenidos mediante los otros dos métodos mencionados en el párrafo anterior. Sólo en algunos estados como en Baja California, Tlaxcala y Zacatecas este índice fue mayor. Respecto a la estimación del error estándar por entidad federativa, se aprecia en las gráficas 2.3 y 2.4 y en la tabla 2.5 que el estimador fue mayor en casi todos los casos, comparado con los obtenidos por los otros métodos.

La estimación puntual de algunas entidades federativas, obtenidas por este método, resultaron ser muy distintas de las obtenidas al considerar las expresiones analíticas de Sandström aplicadas a un diseño muestral aleatorio simple. En la gráfica 2.3 se aprecia que los intervalos de confianza obtenidos por ambos criterios para las entidades de Chiapas, Durango, Hidalgo, Jalisco, Nuevo León, Querétaro, San Luis Potosí y Tamaulipas no se traslapan, lo cual indica que al momento de considerar un diseño muestral a los datos, o los factores de expansión obtenidos a través del diseño muestral, los estimadores cambian considerablemente.

En la tabla 2.4 se muestran los resultados del estimador del índice de Gini propuesto por Sandström (mostrados también en la tabla 2.3) pero con la estimación de la varianza obtenida con la técnica jackknife. En algunas entidades, el estimador de la varianza usando esta técnica fue mayor que el propuesto por Sandström. Los estados en los cuales sucedió esto son: Aguascalientes, Campeche, Coahuila, Guerrero, Estado de México, Michoacán, Morelos, Quintana Roo, Sonora, Tlaxcala, Veracruz y Zacatecas. Para las entidades restantes, el estimador de la varianza por jackknife resultó ser más pequeño que el dado por Sandström. En la gráfica 2.4 se muestran los intervalos de confianza generados por ambos métodos de estimación de varianza.

De igual forma, que en los métodos que consideran muestreo aleatorio simple, no se puede concluir cuál de los métodos de estimación de varianza, cuando se considera diseño muestral complejo, presentados es mejor. Si bien, en algunos casos, se obtuvieron estimadores de varianza más pequeños usando la técnica jackknife, esto no es un indicativo suficiente para determinar que es mejor. Es necesario conocer el verdadero valor de la varianza del índice de Gini para así concluir al respecto

Los cálculos para la estimación de varianza a nivel entidad federativa mediante la técnica jackknife fueron programados en el paquete estadístico **S**-Plus. El inconveniente de este paquete es que al momento de obtener una estimación de la varianza para la muestra a nivel Nacional, el paquete no soportó el tamaño de la base de datos, comprendido por 2'050,294 registros de la muestra total. No fue posible aumentar la memoria temporal para realizar tal proceso. Por esta razón, se usó el paquete **R**, generándose con él el resultado a nivel nacional. Se tuvo que aumentar temporalmente la memoria con que cuenta, ya que la que tiene el paquete por default no era suficiente.

Ahora, considerando los valores estimados del coeficiente de Gini, y observando la gráfica comparativa, los estados de la República Mexicana que pueden considerarse con una distribución de la riqueza más uniforme son: Aguascalientes, 0.4837, Colima, 0.4889, Nayarit, 0.5004, y Sinaloa, 0.5002, ya que su respectivo coeficiente es de los más chicos presentados en las tablas y que están muy cercanos a 0.5. Caso contrario ocurre con los estados de Chiapas, 0.6998, Querétaro, 0.6927, Guerrero, 0.6478, y Oaxaca, 0.6282, que tienen los índices más cercanos a la unidad, lo que indica que la distribucion de la riqueza se aleja demasiado de la equidad.

En este capítulo se revisaron algunos métodos para obtener la estimación de la varianza del índice de Gini. La comparación de la estimación del error estándar de los métodos mencionados se muestra a detalle en la tabla 2.5, por entidad federativa y para el nacional.El primero que se mencionó es el propuesto por Ogwang, el cual no considera un diseño muestral para estimar el índice de Gini y realiza la estimación de la varianza usando la técnica jackknife. Los resultados obtenidos usando el método de Ogwang fueron muy parecidos a los obtenidos usando el método propuesto por Sandström pero considerando un diseño muestral aleatorio simple. La ventaja que presenta el método de Ogwang es que la forma de estimación que propone no requiere de paquetes estadísticos complejos, ya que sólo considera que la base de datos está ordenada por la variable de estudio, en este caso el ingreso mensual promedio, y que se tiene una variable adicional donde se acumulan los valores de la variable de estudio registro a registro. De igual forma, para la estimación de la varianza lo hace mediante el uso de la técnica jackknife, pero propone un algoritmo que facilita en forma considerable los cálculos de esta técnica.

Otro método de estimación del índice de Gini mencionado es el propuesto por Sandström et al. Con este método se consideran los factores de expansión del diseño muestral. Para el estimador puntual del índice de Gini, propuesto por Sandström, se obtuvo la estimación de la varianza utilizando la técnica jackknife. Lo resultados obtenidos por ambos métodos de estimación no fueron similares en varios casos, ya que para algunas entidades federativas la varianza obtenida con la técnica jackknife fue menor, mientras que para otra fue menor con el método de Sandström. Es difícil determinar cuál de los dos métodos es mejor, ya que no se tiene un punto de referencia con el cual comparar los resultados. La ventaja que presenta el usar la técnica jackknife es que es de más fácil manejo y se obtienen resultados en menor tiempo, sólo hay que considerar que abusar de esta técnica de estimación puede llevar a resultados no del todo aceptables ya que se pueden tener complicaciones de estimación, dependiendo del diseño muestral considerado. Si se tuviera que sistematizar este proceso, se recomendaría emplear el método de Sandström, ya que se ha validado la eficacia del mismo y no considera supuestos adicionales que se deben cuidar al usar la técnica jackknife, como son: que el diseño muestral sea estratificado o en varias etapas, de esta manera, se asegura la eficacia de los resultados obtenidos.

Así mismo, se observó que el método de Sandström aplicado a un diseño muestral aleatorio simple da resultados similares a los del método de Ogwang, por lo que emplear el método de Sandström lleva a usar uno que requeriría de sistematización adicional, lo cual no es recomendable. Aunque se aconseja usar el método de Ogwang en el caso en que se desee tener resultados rápidos, ya que usar el de Sandström podría ser mucho más lento, aunque será recomendable.

Para determinar la eficacia de los métodos aquí expuestos, bastaría con conocer la varianza del índice de Gini, desglosada por entidad federativa y para el nacional, y así poder compararla con cada método aquí presentado y definir cuál resultó mejor.

Tabla 2.1: Estimación del índice de Gini, de su error estándar e intervalo de confianza al 95 % mediante el método propuesto por Ogwang, para información de la muestra del XII Censo General de Población y Vivienda 2000 desagregado por Entidad Federativa de la variable Ingreso Total Promedio por Hogares.

ESTADO	**TAMAÑO MUESTRAL**	**ÍNDICE GINI**	**ERROR EST.**	**LÍMITE INF.**	**LÍMITE SUP.**
AGS	17,698	0.4869	0.0109	0.4656	0.5082
BC	33,910	0.5827	0.0089	0.5653	0.6002
BCS	8,610	0.6093	0.0176	0.5748	0.6437
CAM	14,500	0.5942	0.0128	0.5692	0.6193
COAH	44,269	0.5629	0.0080	0.5472	0.5786
COL	11,848	0.5135	0.0127	0.4887	0.5384
CHIS	80,681	0.7485	0.0077	0.7334	0.7637
CHIH	64,017	0.6370	0.0074	0.6224	0.6516
DF	180,982	0.5536	0.0022	0.5492	0.5579
DGO	30,475	0.6579	0.0113	0.6357	0.6800
GTO	70,668	0.6248	0.0079	0.6092	0.6404
GRO	62,615	0.6706	0.0094	0.6521	0.6891
HGO	55,532	0.6272	0.0102	0.6073	0.6471
JAL	132,796	0.5529	0.0044	0.5443	0.5615
MEX	231,073	0.5796	0.0036	0.5725	0.5868
MICH	88,319	0.6146	0.0078	0.5993	0.6299
MOR	33,339	0.5515	0.0105	0.5309	0.5721
NAY	17,144	0.5270	0.0115	0.5045	0.5494
NLN	76,400	0.5716	0.0044	0.5631	0.5802
OAX	125,936	0.6813	0.0060	0.6696	0.6930
PUE	115,994	0.6432	0.0068	0.6298	0.6565
QRO	25,375	0.7367	0.0094	0.7183	0.7552
QTR	15,525	0.5703	0.0095	0.5517	0.5890
SLP	51,320	0.6049	0.0075	0.5903	0.6196
SIN	38,462	0.5163	0.0085	0.4997	0.5329
SON	53,960	0.5525	0.0080	0.5369	0.5682
TAB	35,444	0.5815	0.0077	0.5665	0.5965
TAM	52,717	0.5970	0.0071	0.5832	0.6108
TLA	27,518	0.5265	0.0127	0.5015	0.5514
VER	166,179	0.5791	0.0045	0.5703	0.5879
YUC	49,913	0.5943	0.0100	0.5746	0.6140
ZAC	37,075	0.6253	0.0097	0.6063	0.6443
NACIONAL	**2,050,294**	**0.6243**	**0.0012**	**0.6219**	**0.6266**

Tabla 2.2: Estimación del índice de Gini, de su error estándar e intervalo de confianza al 95 % mediante el método propuesto por Sandström considerando un diseño muestral aleatorio simple, para información de la muestra del XII Censo General de Población y Vivienda 2000 desagregado por Entidad Federativa de la variable Ingreso Total Promedio por Hogares.

ESTADO	**TAMAÑO MUEST.**	**ÍNDICE GINI**	**ERROR EST.**	**LÍMITE INF.**	**SUP.**
AGS	17,698	0.4870	0.0103	0.4669	0.5071
BC	33,910	0.5828	0.0086	0.5659	0.5997
BCS	8,610	0.6095	0.0166	0.5769	0.6421
CAM	14,500	0.5944	0.0120	0.5708	0.6179
COAH	44,269	0.5629	0.0077	0.5479	0.5780
COL	11,848	0.5137	0.0119	0.4903	0.5371
CHIS	80,681	0.7486	0.0073	0.7343	0.7628
CHIH	64,017	0.637	0.0071	0.6231	0.6509
DF	180,982	0.5536	0.0021	0.5495	0.5577
DGO	30,475	0.658	0.0106	0.6371	0.6788
GTO	70,668	0.6248	0.0076	0.6099	0.6397
GRO	62,615	0.6706	0.0089	0.6533	0.688
HGO	55,532	0.6272	0.0095	0.6086	0.6458
JAL	132,796	0.5529	0.0042	0.5447	0.5611
MEX	231,073	0.5796	0.0035	0.5728	0.5865
MICH	88,319	0.6146	0.0073	0.6002	0.6290
MOR	33,339	0.5516	0.0099	0.5321	0.5710
NAY	17,144	0.5271	0.0108	0.5059	0.5483
NLN	76,400	0.5716	0.0042	0.5635	0.5798
OAX	125,936	0.6813	0.0053	0.6709	0.6917
PUE	115,994	0.6432	0.0064	0.6307	0.6557
QRO	25,375	0.7368	0.0089	0.7193	0.7543
QTR	15,525	0.5705	0.0091	0.5527	0.5882
SLP	51,320	0.605	0.0070	0.5913	0.6187
SIN	38,462	0.5164	0.0081	0.5004	0.5323
SON	53,960	0.5526	0.0075	0.5378	0.5673
TAB	35,444	0.5816	0.0072	0.5674	0.5958
TAM	52,717	0.5970	0.0067	0.5838	0.6103
TLA	27,518	0.5265	0.0117	0.5037	0.5494
VER	166,179	0.5791	0.0042	0.5709	0.5874
YUC	49,913	0.5943	0.0093	0.5761	0.6125
ZAC	37,075	0.6253	0.0089	0.6078	0.6428
NACIONAL	**2,050,294**	**0.6243**	**0.0011**	**0.6221**	**0.6265**

Tabla 2.3: Estimación del índice de Gini, de su error estándar e intervalo de confianza al 95 % mediante el método propuesto por Sandström considerando diseño muestral, para información de la muestra del XII Censo General de Población y Vivienda 2000 desagregado por Entidad Federativa de la variable Ingreso Total Promedio por Hogares.

ESTADO	**TAMAÑO MUEST.**	**ÍNDICE GINI**	**ERROR EST.**	**LÍMITE INF.**	**LÍMITE SUP.**
AGS	17,698	0.4837	0.0116	0.4609	0.5064
BC	33,910	0.5949	0.0098	0.5758	0.6141
BCS	8,610	0.5561	0.0197	0.5175	0.5947
CAM	14,500	0.5759	0.0143	0.5478	0.6039
COAH	44,269	0.5483	0.0091	0.5304	0.5662
COL	11,848	0.4889	0.0108	0.4677	0.5101
CHIS	80,681	0.6998	0.0101	0.6799	0.7197
CHIH	64,017	0.6017	0.0098	0.5824	0.6209
DF	180,982	0.5452	0.0023	0.5406	0.5497
DGO	30,475	0.5522	0.0122	0.5283	0.5761
GTO	70,668	0.6093	0.0092	0.5912	0.6274
GRO	62,615	0.6478	0.0122	0.6239	0.6717
HGO	55,532	0.5690	0.0088	0.5517	0.5864
JAL	132,796	0.5242	0.0046	0.5151	0.5332
MEX	231,073	0.5680	0.0044	0.5593	0.5767
MICH	88,319	0.6026	0.0102	0.5827	0.6226
MOR	33,339	0.5418	0.0128	0.5168	0.5669
NAY	17,144	0.5004	0.0117	0.4775	0.5232
NLN	76,400	0.5409	0.0050	0.5311	0.5507
OAX	125,936	0.6282	0.0071	0.6144	0.6421
PUE	115,994	0.5934	0.0085	0.5768	0.6100
QRO	25,375	0.6927	0.0120	0.6692	0.7161
QTR	15,525	0.5415	0.0138	0.5146	0.5685
SLP	51,320	0.5667	0.0081	0.5509	0.5825
SIN	38,462	0.5002	0.0080	0.4845	0.5160
SON	53,960	0.5488	0.0107	0.5279	0.5697
TAB	35,444	0.5818	0.0088	0.5646	0.5991
TAM	52,717	0.5622	0.0080	0.5466	0.5779
TLA	27,518	0.5377	0.0188	0.5009	0.5746
VER	166,179	0.5651	0.0050	0.5552	0.5749
YUC	49,913	0.5689	0.0096	0.5501	0.5878
ZAC	37,075	0.6300	0.0120	0.6065	0.6535
NACIONAL	**2,050,294**	**0.5913**	**0.0015**	**0.5884**	**0.5941**

Tabla 2.4: Estimación del índice de Gini mediante el método propuesto por Sandström, de su error estándar por la técnica jackknife e intervalo de confianza al 95 % considerando diseño muestral, para información de la muestra del XII Censo General de Población y Vivienda 2000 desagregado por Entidad Federativa de la variable Ingreso Total Promedio por Hogares.

ESTADO	**TAMAÑO MUEST.**	**ÍNDICE GINI**	**ERROR EST.**	**LÍMITE INF.**	**SUP.**
AGS	17,698	0.4837	0.0130	0.4581	0.5092
BC	33,910	0.5949	0.0097	0.5758	0.614
BCS	8,610	0.5561	0.0161	0.5245	0.5877
CAM	14,500	0.5759	0.0145	0.5475	0.6043
COAH	44,269	0.5483	0.0093	0.5300	0.5666
COL	11,848	0.4889	0.0090	0.4712	0.5066
CHIS	80,681	0.6998	0.0086	0.6829	0.7167
CHIH	64,017	0.6017	0.0091	0.5838	0.6195
DF	180,982	0.5452	0.0023	0.5407	0.5496
DGO	30,475	0.5522	0.0091	0.5344	0.5700
GTO	70,668	0.6093	0.0086	0.5924	0.6262
GRO	62,615	0.6478	0.0144	0.6196	0.676
HGO	55,532	0.5690	0.0067	0.5559	0.5822
JAL	132,796	0.5242	0.0045	0.5154	0.5329
MEX	231,073	0.5680	0.0046	0.559	0.5771
MICH	88,319	0.6026	0.0114	0.5804	0.6249
MOR	33,339	0.5418	0.0133	0.5157	0.568
NAY	17,144	0.5004	0.0091	0.4825	0.5183
NLN	76,400	0.5409	0.0050	0.5311	0.5507
OAX	125,936	0.6282	0.0063	0.6159	0.6405
PUE	115,994	0.5934	0.0073	0.5791	0.6077
QRO	25,375	0.6927	0.0094	0.6742	0.7111
QTR	15,525	0.5415	0.0154	0.5114	0.5717
SLP	51,320	0.5667	0.0068	0.5534	0.5800
SIN	38,462	0.5002	0.0066	0.4873	0.5132
SON	53,960	0.5488	0.0107	0.5278	0.5698
TAB	35,444	0.5818	0.0084	0.5654	0.5983
TAM	52,717	0.5622	0.0077	0.5472	0.5773
TLA	27,518	0.5377	0.0237	0.4912	0.5843
VER	166,179	0.5651	0.0057	0.5540	0.5762
YUC	49,913	0.5689	0.0083	0.5527	0.5852
ZAC	37,075	0.6300	0.0133	0.6040	0.6560
NACIONAL	**2,050,294**	**0.5913**	**0.0015**	**0.5884**	**0.5941**

Tabla 2.5: Estimación del error estándar, por 100, obtenido mediante los métodos propuestos por: **A.** Ogwang (2000), **B.** Sandström et al. (1988), considerando un m.a.s. en las expresiones dadas, **C.** Sandström et al. (1988) y **D.** técnica jackknife aplicada al estimador del índice de Gini propuesto por Sandström (1988). Cociente de la estimación del error estándar de los métodos mencionados.

ESTADO	**MUESTRA**	**A**	**B**	**C**	**D**	**A/B**	**D/C**	**C/B**
AGS	17,698	1.09	1.03	1.16	1.30	1.06	1.12	1.13
BC	33,910	0.89	0.86	0.98	0.97	1.03	1.00	1.13
BCS	8,610	1.76	1.66	1.97	1.61	1.06	0.82	1.18
CAM	14,500	1.28	1.20	1.43	1.45	1.06	1.01	1.19
COAH	44,269	0.80	0.77	0.91	0.93	1.05	1.02	1.19
COL	11,848	1.27	1.19	1.08	0.90	1.06	0.83	0.91
CHIS	80,681	0.77	0.73	1.01	0.86	1.07	0.85	1.40
CHIH	64,017	0.74	0.71	0.98	0.91	1.05	0.93	1.39
DF	180,982	0.22	0.21	0.23	0.23	1.05	0.98	1.11
DGO	30,475	1.13	1.06	1.22	0.91	1.06	0.74	1.15
GTO	70,668	0.79	0.76	0.92	0.86	1.05	0.93	1.22
GRO	62,615	0.94	0.89	1.22	1.44	1.06	1.18	1.38
HGO	55,532	1.02	0.95	0.88	0.67	1.07	0.76	0.93
JAL	132,796	0.44	0.42	0.46	0.45	1.05	0.97	1.11
MEX	231,073	0.36	0.35	0.44	0.46	1.05	1.04	1.27
MICH	88,319	0.78	0.73	1.02	1.14	1.07	1.12	1.38
MOR	33,339	1.05	0.99	1.28	1.33	1.06	1.04	1.29
NAY	17,144	1.15	1.08	1.17	0.91	1.06	0.78	1.08
NLN	76,400	0.44	0.42	0.50	0.50	1.05	1.00	1.21
OAX	125,936	0.60	0.53	0.71	0.63	1.12	0.89	1.33
PUE	115,994	0.68	0.64	0.85	0.73	1.07	0.86	1.33
QRO	25,375	0.94	0.89	1.20	0.94	1.05	0.79	1.34
QTR	15,525	0.95	0.91	1.38	1.54	1.05	1.12	1.52
SLP	51,320	0.75	0.70	0.81	0.68	1.07	0.84	1.15
SIN	38,462	0.85	0.81	0.80	0.66	1.04	0.82	0.99
SON	53,960	0.80	0.75	1.07	1.07	1.06	1.00	1.42
TAB	35,444	0.77	0.72	0.88	0.84	1.06	0.95	1.22
TAM	52,717	0.71	0.67	0.80	0.77	1.05	0.97	1.18
TLA	27,518	1.27	1.17	1.88	2.37	1.09	1.26	1.61
VER	166,179	0.45	0.42	0.50	0.57	1.06	1.13	1.19
YUC	49,913	1.00	0.93	0.96	0.83	1.08	0.86	1.04
ZAC	37,075	0.97	0.89	1.20	1.33	1.09	1.11	1.34
NACIONAL	**2,050,294**	**0.12**	**0.11**	**0.15**	**0.15**	**1.05**	**1.00**	**1.28**

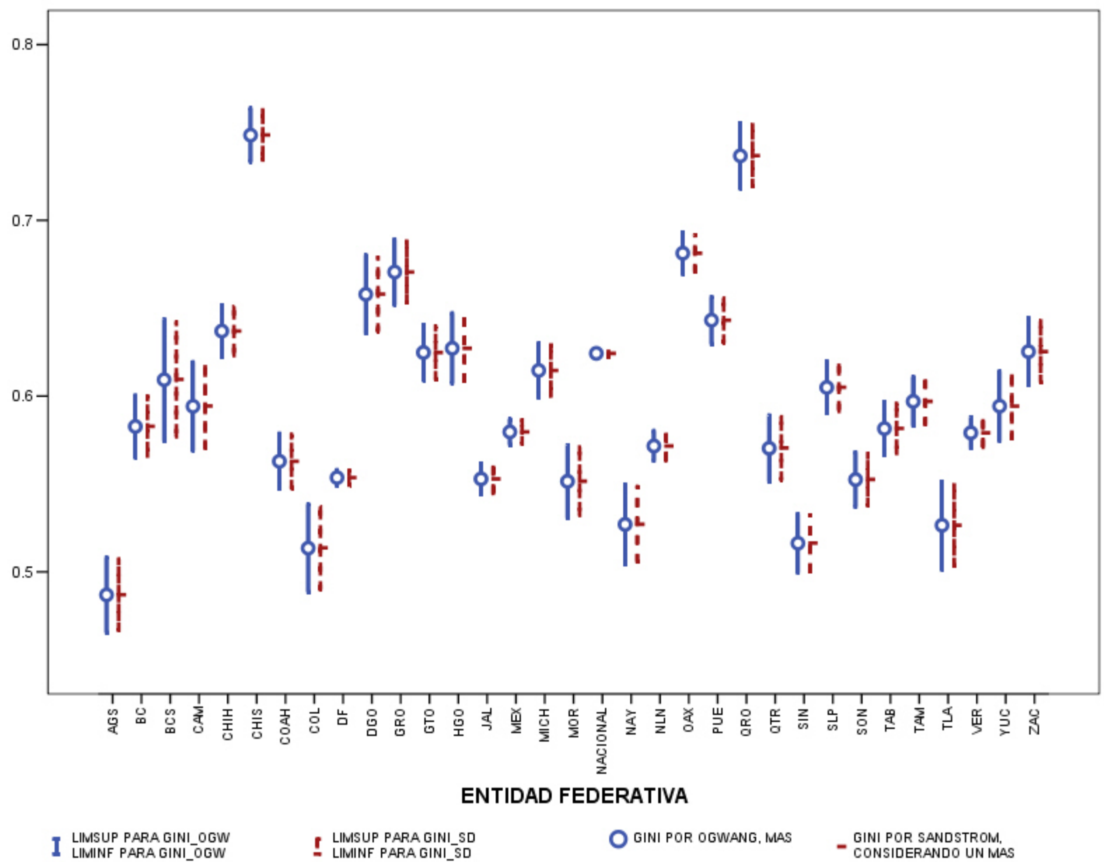

Figura 2.2: Estimación del índice de Gini y de su intervalo de confianza al 95 % (GINI-OGW) por el método propuesto por Ogwang (2000), que utiliza Mínimos Cuadrados Ponderados para la estimación puntual y la técnica jackknife para la estimación de la varianza. Además, índice de Gini e intervalo de confianza al 95 % (GINI-SD), empleando el método propuesto por Sandström (1988) pero considerando diseño muestral aleatorio simple (MAS).

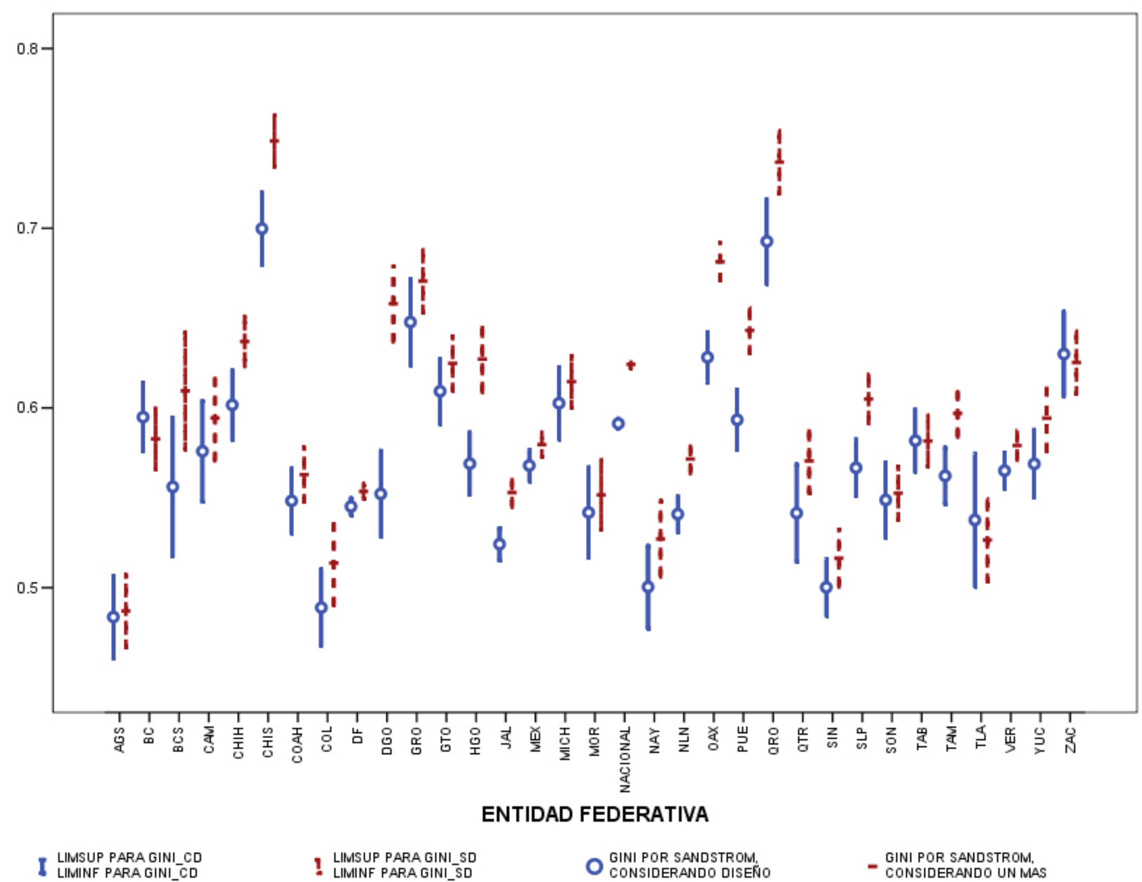

Figura 2.3: Estimación del índice de Gini y de su intervalo del confianza al 95 % por el método propuesto por Sandström (1988), considerando diseño muestral (GINI-CD) y suponiendo un MAS (GINI-SD).

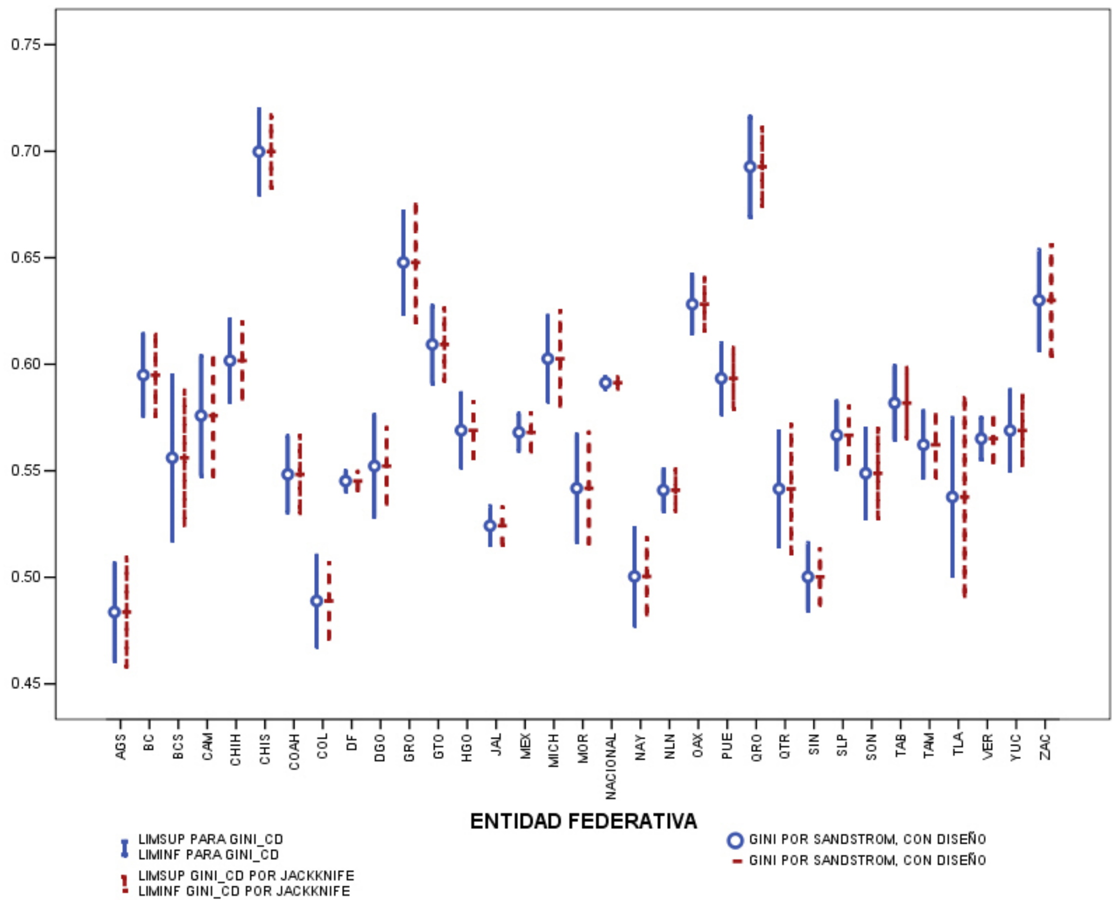

Figura 2.4: Estimación del índice de Gini y de su intervalo de confianza al 95 % por el método propuesto por Sandström (1988) y utilizando la técnica jackknife para la estimación de la varianza. Ambos métodos consideran el diseño muestral en la estimación.

Capítulo 3

Indices Basados en Estimadores de Razón

Introducción

Cuando se realiza la estimación de un índice, y en general de cualquier parámetro, usualmente se desea saber cuál es la variabilidad que presenta tal estimación y así determinar si es confiable o no. Por consiguiente, se desea obtener un intervalo de confianza del parámetro de estudio.

La forma de construir un índice varía, dependiendo de la finalidad del mismo y de los datos disponibles. En el caso de México, se obtienen diversos índices para diferentes áreas de estudio, tales como el Indice Nacional de Precios al Consumidor (INPC), el Indice Nacional de Precios al Productor (INPP), véase [16], el Indice de Marginalidad, el Indice de Precios y Cotizaciones (IPC), entre otros. Cuando se publica la estimación de estos índices, no siempre se acompaña del error de estimación o de un intervalo de confianza. Esto en parte por la dificultad de cálculo que algunos índices presentan en su estimación, haciendo que la estimación de su varianza sea compleja de obtener y que no cuente con una expresión explícita, por consiguiente, la dificultad para presentar un intervalo de confianza.

En esta sección se presenta un estimador de índices basado en estimadores de razón, propuesto por Kish (1968), el cual se ilustrará mediante un caso práctico. Los datos usados serán el Gasto Mensual Promedio en Hogares mexicanos para los años 2002 y 2004, tomando como base el año 2000, de la Encuesta Nacional de Ingresos y Gastos en Hogares (ENIGH). De igual forma, se presenta un estimador de la varianza de estos índices, basado en la aproximación de series de Taylor para un estimador de varianza de estimadores de razón. Como caso alterno, se presentan resultados de estimación de varianza obtenidos mediante la técnica jackknife, visto en la sección §2.4.

3.1. Indice propuesto por Kish

A continuación, se presenta una expresión del estimador de un índice propuesto por Kish (1968). Este estimador tiene su base en el cociente de estimadores de razón para un año en curso (o de un periodo determinado considerado) con un año base (inicial), por lo que se define primero el estimador de razón como:

$$\hat{r} = \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}} = \frac{\sum_s \frac{y_k}{\pi_k}}{\sum_s \frac{x_k}{\pi_k}}. \tag{3.1}$$

donde, tanto el numerador de la expresión como el denominador representan, en este caso, el estimador π del total de las variables de estudio y y x, respectivamente, donde ambas variables pueden pertenecer a información de la misma encuesta (o año de estudio) o ser de diferente origen (año en curso comparado con un año base distinto). En Särndal et al. (1992) se define con más detalle el estimador π y algunas de sus propiedades estadísticas

A partir de este estimador de razón, se define una razón doble, que es el cociente de dos estimadores de razón. A esta expresión Kish le llama *relativa*, "**relative**"

$$\hat{R}_1 = \frac{\hat{r}_1}{\hat{r}_0} = \left(\frac{\hat{t}_{y_1}}{\hat{t}_{x_1}}\right) \Big/ \left(\frac{\hat{t}_{y_0}}{\hat{t}_{x_0}}\right) = \frac{\hat{t}_{y_1}\hat{t}_{x_0}}{\hat{t}_{y_0}\hat{t}_{x_1}} \tag{3.2}$$

donde

$$\hat{r}_1 = \frac{\hat{t}_{y_1}}{\hat{t}_{x_1}}, \qquad \hat{r}_0 = \frac{\hat{t}_{y_0}}{\hat{t}_{x_0}} \tag{3.3}$$

son los estimadores de razón del año en curso (o del periodo de interés), $\hat{r}_1$, y el del año base (o inicial), $\hat{r}_0$, con el cual se desea comparar el cambio.

La estimación de la varianza para estos estimadores de razón está dada por:

$$\widehat{V}(\hat{r}) = \frac{1}{\hat{t}_x^2}\left[\widehat{V}\left(\hat{t}_y\right) + \hat{r}^2\widehat{V}\left(\hat{t}_x\right) - 2\hat{r}\widehat{C}\left(\hat{t}_x, \hat{t}_y\right)\right] \tag{3.4}$$

La aproximación de la varianza de (3.2), considerando que $\hat{R}_1$ es la razón de dos variables aleatorias, está dada por la siguiente expresión, la cual es una aproximación hecha mediante expansión de series de Taylor,

$$\begin{aligned}\widehat{V}\left(\hat{R}_1\right) &= \frac{1}{\hat{r}_0^2}\left[\widehat{V}(\hat{r}_1) + \hat{R}_1^2\widehat{V}(\hat{r}_0) - 2\hat{R}_1\widehat{C}(\hat{r}_0, \hat{r}_1)\right] \\ &= \hat{R}_1^2\left[\frac{\widehat{V}(\hat{r}_1)}{\hat{r}_1} + \frac{\widehat{V}(\hat{r}_0)}{\hat{r}_0^2} - 2\frac{\widehat{C}(\hat{r}_0, \hat{r}_1)}{\hat{r}_0\hat{r}_1}\right] \end{aligned} \tag{3.5}$$

En Kish (1979) se presenta la siguiente expresión de la covarianza entre dos estimadores de razón, $\hat{r}_0$ y $\hat{r}_1$, la cual depende únicamente de la estimación de covarianzas de estimadores π de totales,

$$\widehat{C}\left(\hat{r}_0, \hat{r}_1\right) = \frac{1}{\hat{t}_{x_0}\hat{t}_{x_1}}\left[\widehat{C}\left(\hat{t}_{y_0}, \hat{t}_{y_1}\right) + \hat{r}_0\hat{r}_1\widehat{C}\left(\hat{t}_{x_0}, \hat{t}_{x_1}\right) - \hat{r}_0\widehat{C}\left(\hat{t}_{y_0}, \hat{t}_{x_1}\right) - \hat{r}_1\widehat{C}\left(\hat{t}_{y_1}, \hat{t}_{x_0}\right)\right] \tag{3.6}$$

donde $\hat{r}_0$ y $\hat{r}_1$ se definen como en (3.3).

Usando las expresiones (3.4) y (3.6), la expresión (3.5) se convierte en:

$$\begin{aligned}
\widehat{V}\left(\hat{R}_1\right) &= \left(\frac{\hat{t}_{x_0}}{\hat{t}_{y_0}}\right)^2 \left\{\frac{1}{\hat{t}_{x_1}^2}\left[\widehat{V}\left(\hat{t}_{y_1}\right) + \hat{r}_1\widehat{V}\left(\hat{t}_{x_1}\right) - 2\hat{r}_1\widehat{C}\left(\hat{t}_{x_1}, \hat{t}_{y_1}\right)\right]\right\} \\
&+ \left(\frac{\hat{t}_{y_1}\hat{t}_{x_0}}{\hat{t}_{y_0}\hat{t}_{x_1}}\right)^2 \left\{\frac{1}{\hat{t}_{y_0}^2}\left[\widehat{V}\left(\hat{t}_{y_0}\right) + \hat{r}_0\widehat{V}\left(\hat{t}_{x_0}\right) - 2\hat{r}_0\widehat{C}\left(\hat{t}_{x_0}, \hat{t}_{y_0}\right)\right]\right\} \\
&-2\left(\frac{\hat{t}_{y_1}\hat{t}_{x_0}}{\hat{t}_{y_0}\hat{t}_{x_1}}\right) \left\{\frac{1}{\hat{t}_{x_0}\hat{t}_{x_1}}\left[\widehat{C}\left(\hat{t}_{y_1}, \hat{t}_{y_0}\right) + \hat{r}_0\hat{r}_1\widehat{C}\left(\hat{t}_{x_0}, \hat{t}_{x_1}\right)\right]\right\} \\
&+2\left(\frac{\hat{t}_{y_1}\hat{t}_{x_0}}{\hat{t}_{y_0}\hat{t}_{x_1}}\right) \left\{\frac{1}{\hat{t}_{x_1}\hat{t}_{x_0}}\left(\frac{\hat{t}_{x_0}^2}{\hat{t}_{y_0}^2}\right)\left[\hat{r}_0\widehat{C}\left(\hat{t}_{x_1}, \hat{t}_{y_0}\right) + \hat{r}_1\widehat{C}\left(\hat{t}_{x_0}, \hat{t}_{y_1}\right)\right]\right\}.
\end{aligned}$$

Que es una expresión más extensa que la mostrada en (3.5) pero en cuanto a su cálculo resulta ser más sencilla, debido a que sólo depende de la estimación π de los totales de las variables consideradas para los años respectivos. El inconveniente que muestra esta expresión, es que algunas de las covarianzas presentan combinación de estimadores π de totales de diferentes encuestas (años), tales como $\widehat{C}\left(\hat{t}_{y_1}, \hat{t}_{y_0}\right)$, $\widehat{C}\left(\hat{t}_{x_0}, \hat{t}_{x_1}\right)$, $\widehat{C}\left(\hat{t}_{x_1}, \hat{t}_{y_0}\right)$ y $\widehat{C}\left(\hat{t}_{x_0}, \hat{t}_{y_1}\right)$, por lo que no se puede obtener su estimación en forma directa o con algún paquete estadístico, debido a que estos no consideran fuentes de información de distintas encuestas en el proceso de estimación. Si las encuestas son independientes, entonces las covarianzas mencionadas son cero.

Si ese fuera el caso, la estimación de $\widehat{V}\left(\hat{R}_1\right)$ se reduce a

$$\begin{aligned}
\widehat{V}\left(\hat{R}_1\right) &= \left(\frac{\hat{t}_{x_0}}{\hat{t}_{y_0}}\right)^2 \left\{\frac{1}{\hat{t}_{x_1}^2}\left[\widehat{V}\left(\hat{t}_{y_1}\right) + \hat{r}_1\widehat{V}\left(\hat{t}_{x_1}\right) - 2\hat{r}_1\widehat{C}\left(\hat{t}_{x_1}, \hat{t}_{y_1}\right)\right]\right\} \\
&+ \left(\frac{\hat{t}_{y_1}\hat{t}_{x_0}}{\hat{t}_{y_0}\hat{t}_{x_1}}\right)^2 \left\{\frac{1}{\hat{t}_{y_0}^2}\left[\widehat{V}\left(\hat{t}_{y_0}\right) + \hat{r}_0\widehat{V}\left(\hat{t}_{x_0}\right) - 2\hat{r}_0\widehat{C}\left(\hat{t}_{x_0}, \hat{t}_{y_0}\right)\right]\right\}.
\end{aligned}$$

Una aproximación a la expresión (3.5) es la dada por Kish (1968) en su artículo, la cual tiene las siguientes alternativas de cálculo.

$$
\begin{aligned}
\widehat{V}\left(\hat{R}_1\right) &= \hat{R}_1^2 \sum_{h=1}^{H}\left[\left(\frac{d\hat{t}_{y_1}}{\hat{t}_{y_1}}-\frac{d\hat{t}_{x_1}}{\hat{t}_{x_1}}\right)-\left(\frac{d\hat{t}_{y_0}}{\hat{t}_{y_0}}-\frac{d\hat{t}_{x_0}}{\hat{t}_{x_0}}\right)\right]_h^2 \\
&\doteq \frac{1}{\hat{r}_0^2}\sum_{h=1}^{H}\left(\frac{d\hat{t}_{z_{1h}}}{\hat{t}_{x_1}}-\hat{R}_1\frac{d\hat{t}_{z_{0h}}}{\hat{t}_{x_0}}\right)^2 \\
&= \sum_{h=1}^{H}\left(e_{1h}-\hat{R}_1 e_{0h}\right)^2 \\
&= \sum_{h=1}^{H} e_{1h}^2 - 2\hat{R}_1\sum_{h=1}^{H} e_{0h}e_{1h} + \hat{R}_1^2\sum_{h=1}^{H} e_{0h}^2 \qquad (3.7)
\end{aligned}
$$

donde $e_{1h} = \frac{d\hat{t}_{z_{1h}}}{\hat{r}_0\hat{t}_{x_1}} = \frac{d\hat{t}_{y_{1h}}-\hat{r}_1 d\hat{t}_{x_{1h}}}{\hat{r}_0\hat{t}_{x_1}}$, $e_{0h} = \frac{d\hat{t}_{z_{0h}}}{\hat{t}_{y_0}} = \frac{d\hat{t}_{y_{0h}}-\hat{r}_0 d\hat{t}_{x_{0h}}}{\hat{t}_{y_0}}$ y $d\hat{t}_{y_{1h}} = \left(\hat{t}_{y_{1h}}-\frac{\hat{t}_{y_{1h}}}{n_{1h}}\right)$, $d\hat{t}_{x_{0h}} = \left(\hat{t}_{x_{0h}}-\frac{\hat{t}_{x_{0h}}}{n_{0h}}\right)$

Tal aproximación de la varianza se basa en un diseño muestral estratificado, con H estratos primarios y n_h es el número de UPM's para el $h-\textit{ésimo}$ estrato.

Haciendo analogía con la ecuación (3.5) se tiene lo siguiente

$$
\begin{aligned}
\sum_{h=1}^{H} e_{1h}^2 &= \frac{\widehat{V}\left(\hat{r}_1\right)}{\hat{r}_0^2} \qquad (3.8)\\
\hat{R}_1^2\sum_{h=1}^{H} e_{0h}^2 &= \hat{R}_1^2\frac{\widehat{V}\left(\hat{r}_0\right)}{\hat{r}_0^2} \qquad (3.9)\\
-2\hat{R}_1\sum_{h=1}^{H} e_{0h}e_{1h} &= -2\hat{R}_1\frac{\widehat{C}\left(\hat{r}_0,\hat{r}_1\right)}{\hat{r}_0^2}. \qquad (3.10)
\end{aligned}
$$

Como se mencionó anteriormente, la covarianza entre dos estimadores de razón, $\widehat{C}\left(\hat{r}_0,\hat{r}_1\right)$, en este caso tiene el inconveniente de considerar información de dos muestras diferentes, esto es, cada uno de los estimadores de razón se obtienen de una fuente de información distinta, por lo que su cálculo es complejo[1]. Por esta razón, la información y los procedimientos de cálculo no se pueden adecuar tan fácil a los estándares de los paquetes estadísticos mencionados al inicio de este trabajo, los

[1]Cuando se consideran encuestas tipo pánel, la covarianza debe estar presente en el cálculo de la estimación de parámetros. En el caso que las encuestas independientes consideradas se levantaran en distintas UPM, entonces se tiene que $\widehat{C}\left(\hat{r}_0,\hat{r}_1\right)=0$.

cuales no consideran la estimación de parámetros cuya expresión mezcle información de dos muestras diferentes. Básicamente, la estimación de estos parámetros se obtuvo programando la metodología de cálculo, la cual resultó un poco engorrosa debido a que previamente se tuvieron que preparar las bases de datos de la ENIGH con la forma adecuada que exige la estimación de varianzas propuesta por Kish.

Cabe mencionar que Kish, (1968) página 513-516, menciona que las expresiones usadas consideran información de encuestas diferentes que están correlacionadas, razón por la cual el término de la covarianza aparece en sus cálculos. Además, menciona que si las encuestas usadas hubieran sido tipo panel, con mismas UPM, entonces la correlación obtenida hubiera sido mayor a la obtenida.

Cochran (1976) presenta una expresión explícita para la estimación de la covarianza entre dos estimadores de razón, para el caso que la información proviene de la misma encuesta. Se utilizó esta expresión como base para obtener la expresión que considera la covarianza de dos parámetros provenientes de muestras diferentes y que no son independientes. De esta manera, considerando un diseño muestral estratificado, la expresión de la covarianza queda de la siguiente manera

$$\widehat{C}\left(\hat{r}_{0j}, \hat{r}_{1k}\right) = \sum_{h=1}^{H} \frac{1}{\hat{t}_{x_{0j}} \hat{t}_{x_{1k}}} \left(\frac{N_{0jh} N_{1kh}}{\sqrt{n_{0jh}\left(n_{0jh}-1\right)}\sqrt{n_{1kh}\left(n_{1kh}-1\right)}} \right) \sum_{l=1}^{n_{0jh}} \sum_{i=1}^{n_{1kh}} d_{0jhl} d_{1khi} \tag{3.11}$$

donde los subíndices j y k representan las categorías (en caso de haber varias) del año indicado ($0-$inicial o base, $1-$de estudio o final) y

$$\begin{aligned} d_{0jhl} &= \left(y_{0jhl} - \bar{y}_{0jh}\right) - \hat{r}_{0j}\left(x_{0jhl} - \bar{x}_{0jh}\right) \\ d_{1khi} &= \left(y_{1khi} - \bar{y}_{1kh}\right) - \hat{r}_{1k}\left(x_{1khi} - \bar{x}_{1kh}\right) \\ \bar{y}_{1kh} &= \frac{1}{N_{1kh}} \sum_{i=1}^{n_{kh}} \frac{y_{1khi}}{\pi_i} \end{aligned}$$

en forma equivalente se obtienen las medias muestrales $\bar{x}_{1kh}$, $\bar{y}_{0jh}$ y $\bar{x}_{0jh}$. La estimación de la expresión dada en (3.11) supone que se conoce el total de unidades de muestreo en la población para cada estrato. Haciendo analogía de estos resultados y aplicándolos a las expresiones dadas en (3.8) (3.9) y (3.10) se tiene que

$$
\begin{aligned}
\sum_{h=1}^{H} e_{1h}^{2} &= \frac{\widehat{V}\left(\hat{r}_1\right)}{\hat{r}_0^2} \\
&= \sum_{h=1}^{H}\left(\frac{1}{\hat{r}_0^2\hat{t}_{x_1}^2}\right)\left(\frac{N_{1h}^2}{n_{1h}\left(n_{1h}-1\right)}\sum_{i=1}^{n_{1h}} d_{1hi}^2\right) \qquad (3.12)
\end{aligned}
$$

$$
\begin{aligned}
\hat{R}_1^2\sum_{h=1}^{H} e_{0h}^{2} &= \hat{R}_1^2\frac{\widehat{V}\left(\hat{r}_0\right)}{\hat{r}_0^2} \\
&= \sum_{h=1}^{H}\hat{R}_1^2\left(\frac{1}{\hat{r}_0^2\hat{t}_{x_0}^2}\right)\left(\frac{N_{0h}^2}{n_{0h}\left(n_{0h}-1\right)}\sum_{j=1}^{n_{0h}} d_{0hj}^2\right) \qquad (3.13)
\end{aligned}
$$

$$
\begin{aligned}
-2\hat{R}_1\sum_{h=1}^{H} e_{0h}e_{1h} &= -2\hat{R}_1\frac{\widehat{C}\left(\hat{r}_0,\hat{r}_1\right)}{\hat{r}_0^2} \\
&= -2\hat{R}_1\frac{1}{\hat{r}_0^2}\sum_{h=1}^{H}\frac{1}{\hat{t}_{x_0}\hat{t}_{x_1}} * \\
&\quad \frac{N_{0h}N_{1h}}{\sqrt{n_{0h}\left(n_{0h}-1\right)}\sqrt{n_{1h}\left(n_{1h}-1\right)}}\sum_{l=1}^{n_{0h}}\sum_{i=1}^{n_{1h}} d_{0hl}d_{1hi} \qquad (3.14)
\end{aligned}
$$

de esta manera, conociendo cada uno de los términos mencionados, es posible obtener la estimación de la varianza de la relativa $\hat{R}_1$, ecuación (3.7), mediante el método propuesto por Kish.

Además de la estimación de varianza para una relativa, también se considera importante la diferencia de dos razones dobles, relativas, para medir el cambio existente en relativas del periodo 1 al periodo 2. Esta diferencia está dada por la siguiente expresión

$$
\begin{aligned}
\widehat{V}\left(\hat{R}_2-\hat{R}_1\right) &= \widehat{V}\left(\hat{R}_1\right)+\widehat{V}\left(\hat{R}_2\right)-2\widehat{C}\left(\hat{R}_1,\hat{R}_2\right) \\
&= \frac{1}{\hat{r}_0^2}\widehat{V}\left(\hat{r}_2-\hat{r}_1\right)+\frac{1}{\hat{r}_0^2}\left(\hat{R}_2-\hat{R}_1\right)^2\widehat{V}\left(\hat{r}_0\right) \\
&\quad -2\frac{1}{\hat{r}_0^2}\left(\hat{R}_2-\hat{R}_1\right)\left[\widehat{C}\left(\hat{r}_0,\hat{r}_2\right)-\widehat{C}\left(\hat{r}_0,\hat{r}_1\right)\right] \qquad (3.15)
\end{aligned}
$$

donde $\hat{R}_1=\frac{\hat{r}_1}{\hat{r}_0}$ y $\hat{R}_2=\frac{\hat{r}_2}{\hat{r}_0}$.

La expresión explícita para obtener la estimación de este parámetro se deriva de las ecuaciones (3.12), (3.13) y (3.14), adecuándolas al año correspondiente.

Si se considera una suma ponderada, $\sum_j^J W_j\hat{R}_{1j}$, donde $\hat{R}_{1j}=\frac{\hat{r}_{1j}}{\hat{r}_0}$, se forma un índice de varias relativas. Nótese que todas las relativas de la suma corresponden al mismo periodo de estudio, lo único que cambia es el subíndice j, que indica la intervención de otra variable de estudio en la relativa correspondiente. La estimación de la varianza de este índice ponderado está dada por

$$\begin{aligned}\widehat{V}\left(\sum_{j=1}^{J} W_j\hat{R}_{1j}\right) &= \sum_{j=1}^{J}\widehat{V}\left(W_j\hat{R}_{1j}\right)+2\sum_{\substack{j=1\\ j\neq k}}^{J}\sum_{k=1}^{J}\widehat{C}\left(W_j\hat{R}_{1j},W_k\hat{R}_{1k}\right)\\ &= \sum_{j=1}^{J}W_j^2\widehat{V}\left(\hat{R}_{1j}\right)+2\sum_{\substack{j=1\\ j\neq k}}^{J}\sum_{k=1}^{J}W_jW_k\widehat{C}\left(\hat{R}_{1j},\hat{R}_{1k}\right) \quad (3.16)\end{aligned}$$

El interés principal en el presente trabajo es sobre el índice $\hat{I}_1=\sum_{j=1}^{J}\hat{R}_{1j}/J$, el cual es un índice no ponderado y cuya varianza se puede obtener en forma explícita de la expresión (3.16), considerando todas las $W_j=1$, y que está dada por

$$\begin{aligned}\widehat{V}\left(\hat{I}_1\right) &= \widehat{V}\left(\frac{1}{J}\sum_{j=1}^{J}\hat{R}_{1j}\right)\\ &= \frac{1}{J^2}\left[\sum_{j=1}^{J}\widehat{V}\left(\hat{R}_{1j}\right)+2\sum_{\substack{j=1\\ j\neq k}}^{J}\sum_{k=1}^{J}\widehat{C}\left(\hat{R}_{1j},\hat{R}_{1k}\right)\right] \quad (3.17)\end{aligned}$$

Aplicando los resultados dados por Kish a esta última expresión, se obtiene una que en lugar de estar en términos de la covarianza de las relativas, $\widehat{C}\left(\hat{R}_{1j},\hat{R}_{1k}\right)$, está en términos de las covarianzas entre estimadores de razón, cuya expresión se mencionó anteriormente.

$$\begin{aligned}\widehat{V}\left(\hat{I}_1\right) &= \frac{1}{J^2}\Bigg[\sum_{j=1}^{J}\widehat{V}\left(\hat{R}_{1j}\right)+2\sum_{\substack{j=1\\ j\neq k}}^{J}\sum_{k=1}^{J}\widehat{C}\left(\hat{r}_{1j},\hat{r}_{1k}\right)\\ &\quad -4\sum_{\substack{j=1\\ j\neq k}}^{J}\sum_{k=1}^{J}\hat{R}_{1j}\widehat{C}\left(\hat{r}_{0j},\hat{r}_{1k}\right)+2\sum_{\substack{j=1\\ j\neq k}}^{J}\sum_{k=1}^{J}\hat{R}_{1j}\hat{R}_{1k}\widehat{C}\left(\hat{r}_{0j},\hat{r}_{0k}\right)\Bigg] \quad (3.18)\end{aligned}$$

La forma de obtener la estimación de esta expresión se deriva de las ecuaciones (3.12), (3.13) y (3.14) para la estimación del primer término, $\widehat{V}\left(\hat{R}_{1j}\right)$, y de la expresión dada en la ecuación (3.11), para la estimación de $\widehat{C}\left(\hat{r}_{1j}, \hat{r}_{1k}\right)$, $\widehat{C}\left(\hat{r}_{0j}, \hat{r}_{1k}\right)$ y $\widehat{C}\left(\hat{r}_{0j}, \hat{r}_{0k}\right)$, considerando el año y las categorías indicadas en los subíndices. En el caso de que las encuestas consideradas sean independientes, el tercer término de la expresión (3.18) desaparece, y en el caso de la expresión $\widehat{V}\left(\hat{R}_{1j}\right)$ la covarianza también es cero.

Finalmente, si se desea la varianza de la diferencia de dos índices $\left[\hat{I}_2 - \hat{I}_1\right]$; ésta se obtiene de (3.16) con $W_{2j} = 1$ y $W_{1j} = -1$, $j = 1, \ldots, J$,. Entonces, el cambio del periodo 1 al periodo 2 de un índice tendrá la varianza

$$\begin{aligned}
\widehat{V}\left(\hat{I}_2 - \hat{I}_1\right) &= \widehat{V}\left[\frac{1}{J}\left(\sum_{j=1}^{J}\hat{R}_{2j} - \sum_{k=1}^{J}\hat{R}_{1k}\right)\right] \\
&= \frac{1}{J^2}\left[\widehat{V}\left(\sum_{j=1}^{J}\hat{R}_{2j}\right) + \widehat{V}\left(\sum_{k=1}^{J}\hat{R}_{1k}\right) - 2\sum_{j=1}^{J}\sum_{k=1}^{J}\widehat{C}\left(\hat{R}_{2j}, \hat{R}_{1k}\right)\right]
\end{aligned}$$

donde $\widehat{V}\left(\sum_{j=1}^{J}\hat{R}_{2j}\right)$ y $\widehat{V}\left(\sum_{k=1}^{J}\hat{R}_{1k}\right)$ se calculan como en (3.17) para el periodo respectivo, mientras que la covarianza entre $\hat{R}_{2j}$ y $\hat{R}_{1k}$ está dada por

$$\widehat{C}\left(\hat{R}_{2j}, \hat{R}_{1k}\right) = \frac{-1}{\hat{r}_0^2}\left[\widehat{C}\left(\hat{r}_{2k}, \hat{r}_{1j}\right) + \hat{R}_{2k}\hat{R}_{1j}\widehat{V}\left(\hat{r}_0\right)\right]$$

donde $\widehat{C}\left(\hat{r}_{2k}, \hat{r}_{1j}\right)$ se calcula de la misma manera que la expresión dada en (3.6).

Una vez obtenidas estas expresiones analíticas, es posible obtener la estimación del índice de Kish y su varianza a cualquier caso práctico en el que sea posible la construcción de estimadores de razón con las características descritas previamente.

3.2. Evaluación del Indice de Kish

El índice propuesto por Kish se usó en el presente trabajo para la construcción de un índice que representa el cambio en el Gasto Mensual Promedio en hogares mexicanos para los años 2002 y 2004, tomando como año base el 2000.

Las bases de datos que se usaron para tal efecto son las correspondientes al gasto en los hogares mexicanos de la Encuesta Nacional de Ingresos y Gastos en Hogares, ENIGH, en sus ediciones 2000, 2002 y 2004. Los diseños muestrales de estas encuestas son polietápicos, estratificados y por conglomerados, donde la unidad última de selección es la vivienda y la unidad de observación es el hogar.

La información que se presenta en estas bases de datos es de difícil manejo, debido a que se tienen cada una de las variables de gastos de productos y servicios realizados en los hogares de las familias mexicanas, sumando un total de más de 800 variables, esto para cada uno de los años considerados. Para simplificar la dificultad del problema, se decidió trabajar con grupos o categorías más generales, cuyas variables de gastos estuvieran relacionadas entre sí [2]. La selección de las variables y su agrupamiento se hizo considerando las categorías que presenta el INPC en sus resultados, obteniéndose así sólo ocho categorías, las cuales se muestran a continuación con algunos de los productos y/o servicios que incluyen. Para más información véase [1] o la página web del Banco de México.

(a) **Alimentos, bebidas y tabaco,** (A) : que incluye todo tipo de

- *i.* Pan, tortillas y cereales;
- *ii.* Carne;
- *iii.* Pescados y mariscos;
- *iv.* Leche, derivados de leche y huevo;
- *v.* Aceites y grasas comestibles;
- *vi.* Frutas y hortalizas;
- *vii.* Azúcar, café y refrescos envasados;
- *viii.* Otros alimentos como condimentos, chocolates, golosinas y alimentos cocinados fuera de casa;
- *ix.* Bebidas alcohólicas;
- *x.* Tabaco;

(b) **Ropa, calzado y accesorios,** (R) : que incluye todo tipo de

- *i.* Ropa de hombre, mujer, niños y bebés;
- *ii.* Ropa de abrigo y uniformes escolares;
- *iii.* Calzado;
- *iv.* Accesorios y cuidados del vestido;

(c) **Vivienda,** (V) : que incluye todo lo referente a

- *i.* Costo de uso de vivienda, alquilada y rentada;

[2]La forma de agrupación considerada corresponde a las categorías que son utilizadas en la construcción del Indice Nacional de Precios al Consumidor.

ii. Electricidad y combustibles;

iii. Otros servicios relacionados con la vivienda;

(d) **Muebles, aparatos y accesorios domésticos,** (M) : todo lo referente a

i. Muebles y aparatos domésticos;

ii. Accesorios y artículos de limpieza para el hogar;

(e) **Salud y cuidado personal,** (S) : todo lo referente a

i. Salud, medicamentos, aparatos y servicios médicos;

ii. Cuidado personal;

(f) **Transporte,** (T) : refiriéndose a

i. Transporte público;

ii. Transporte por cuenta propia;

(g) **Educación y esparcimiento,** (E) : todo lo relacionado a

i. Educación, desde jardín de niños hasta universidad y materiales escolares y libros;

ii. Esparcimiento como cine, servicios turísticos, discos, juguetes, artículos deportivos, internet y otros;

(h) **Otros servicios,** (O) : referentes a

i. Restaurantes, bares y similares;

ii. Servicios profesionales;

iii. Servicios diversos como funerarios, cuotas de licencias y otros documentos.

Esta agrupación se realizó en las bases de datos de gastos en hogares para los tres años mencionados. De esta manera se homogeneizaron las variables, haciendo más sencilla la comparación de resultados para los distintos años de estudio. La letra mayúscula que aparece en paréntesis es una etiqueta que se le dio a cada categoría y que se usará posteriormente en la notación de los estimadores tanto de razón como de relativas.

La idea inicial de este capítulo, fue hacer la construcción del Indice Nacional de Precios al Consumidor (INPC) para, una vez obtenidas las expresiones explícitas del cálculo de tal índice, obtener un estimador de su varianza que vaya acorde a la forma de obtener el estimador puntual. El inconveniente principal que presentó el cálculo de este índice es que requiere varias fuentes de información. La obtenida de la ENIGH no resultó suficiente para los propósitos mencionados.

De la ENIGH sólo se obtiene información referente al Gasto Mensual Promedio en hogares mexicanos de los diferentes productos y servicios de consumo. Adicional a esta información, se requiere saber la variación en el precio, a través del tiempo, que han tenido cada uno de los diversos productos y servicios de consumo en los hogares mexicanos. Esto es, revisar cuánto ha variado el precio del kilo de huevo de distintas marcas, litro de leche de diferentes marcas y presentaciones, tortilla de maíz y harina de trigo, pan, libros, escuela, transporte, etc., respecto al año o punto de referencia con el momento o año en que se obtiene la estimación del INPC.

El punto de referencia, en términos del INPC, es el año base a partir del cual se empieza a generar la variación en el precio de los productos y/o servicios. En este caso, coincide con el año inicial para la construcción del índice de Kish, ya que el último cambio de base del INPC se realizó en el año 2000, véase [1]. Además, se debe obtener información de cuánto ha variado el consumo de cada uno de estos productos en los hogares mexicanos, lo cual no es sencillo. Para obtener tal información se levanta un muestreo de los diferentes productos y marcas en las tiendas de autoservicio. Se revisa cuál es el espacio, en metros cuadrados, que ocupa cada producto en estantería, a partir de esto, se hace una estimación de lo que se vende en un periodo de tiempo determinado.

Por la complejidad que esto lleva, y debido a que la información referente al consumo de los diferentes productos no está disponible, no fue posible la construcción del INPC, además de que el proceso se lleva a cabo para cada producto y servicio de consumo (siendo más de 300 productos agrupados), véase [1], agrupándose así para los niveles mencionados anteriormente, dependiendo el origen del producto, hasta llegar a las ocho categorías arriba mencionadas.

El desglose más general, mencionado arriba con detalle, de los productos y servicios de consumo en los hogares mexicanos genera un manejo de la información más sencillo. Además, el desglose general de las variables de productos y servicios que están en las bases de datos de la encuesta ENIGH, para las ediciones 2000, 2002 y 2004, presentaron algunas diferencias. Para el año 2004, se encontró que se incorporaron nuevos productos a la encuesta, véase [25], con el objeto de adecuarse a los cambios económicos del país y obtener resultados que reflejen la realidad de consumo en los hogares. Tales productos son los relacionados a la telefonía celular, equipos de cómputo y servicios de internet, entre otros, que se han integrado como parte importante del consumo y del gasto en hogares mexicanos. La integración de estos productos generó un problema al momento de comparar la información, en el supuesto que se hiciera en el nivel más fino (por producto y/o servicio). Al generar las categorías más gruesas, se pierden los productos y servicios que se integran año con año a la economía, quedándose al final sólo con variables que incluyen todos los genéricos de características similares. Al tener esta clasificación en cada una de las ENIGH consideradas, fue posible realizar la comparación del gasto mensual promedio en los hogares.

Una vez estandarizada la información de las ENIGH en las categorías mencionadas, se construyeron estimadores de razón para el estimador π del gasto mensual promedio de cada una de ellas, tomando el denominador como el Gasto Mensual Promedio Total (GT) del año respectivo. Se denotó con un subíndice la etiqueta dada a la categoría que corresponde ($A,\ R,\ V,\ M,\ S,\ T,\ E,\ O$). El año de la encuesta de la que proviene la información se marcó sólo con el último dígito de ese año; $2000 - 0$, $2002 - 2$ y $2004 - 4$. Como ejemplo, considérese el caso del estimador de razón de Alimentos, bebidas y tabaco (A) respecto al Gasto Total (GT) para la ENIGH del año 2002, s_2, esto es,

$$\hat{r}_{A2} = \frac{\hat{t}_{A2}}{\hat{t}_{GT2}} = \frac{\sum_{s_2} \frac{A_k}{\pi_k}}{\sum_{s_2} \frac{GT_k}{\pi_k}} \tag{3.19}$$

donde $\hat{t}_{A2}$ y $\hat{t}_{GT2}$ son los estimadores π del total del gasto mensual promedio en Alimentos, bebidas y tabaco y del Gasto Total, respectivamente, para el año 2002. De manera equivalente, se obtienen los estimadores de razón para las siete categorías restantes en los años 2000 y 2004.

También se generaron otros estimadores de razón alternos que fueron de utilidad en la construcción de las relativas. Estos tienen la siguiente forma

$$\hat{r}'_{A2} = \frac{\hat{t}_{A2}}{\hat{t}_{A0}}, \qquad \hat{r}'_{GT2} = \frac{\hat{t}_{GT2}}{\hat{t}_{GT0}} \tag{3.20}$$

que son los estimadores de razón correspondientes a las categorías de Alimentos y Gasto Total, respectivamente. Nótese que tanto el numerador como el denominador pertenecen a la misma variable de estudio, la diferencia con los estimadores de razón presentados en la ecuación (3.19) radica en que el numerador es el estimador π del total en el año 2002, mientras que el denominador es el correspondiente al año 2000 para la misma categoría o variable de estudio. Estos estimadores miden el cambio en el gasto mensual promedio por categoría tomando al año 2000 como base, por lo que aquí ya se mezcla información proveniente de fuentes diferentes.

Ahora, como se mencionó en la sección §3.1, la construcción de la relativa es el cociente de dos estimadores de razón. El año base a considerar es el 2000, por lo que las comparaciones se harán para los años 2002 y 2004. Entonces, la relativa de

Alimentos (A) para el año 2002 es

$$\begin{aligned}\hat{R}_{A2} &= \frac{\hat{r}_{A2}}{\hat{r}_{A0}} = \frac{\frac{\hat{t}_{A2}}{\hat{t}_{GT2}}}{\frac{\hat{t}_{A0}}{\hat{t}_{GT0}}} \\ &= \left(\frac{\hat{t}_{A2}}{\hat{t}_{A0}}\right)\left(\frac{\hat{t}_{GT0}}{\hat{t}_{GT2}}\right) \\ &= \frac{\hat{r}'_{A2}}{\hat{r}'_{GT2}}\end{aligned} \tag{3.21}$$

por lo que es equivalente obtener las relativas usando los estimadores de razón dados en (3.19) o los mostrados en (3.20). En forma análoga, se obtienen las relativas para las categorías restantes del 2002 y de igual forma para todas las categorías del año 2004.

A partir de estas expresiones es posible la construcción del índice $\left(\hat{I}_2\right)$ propuesto por Kish, que no es más que el promedio no ponderado de las relativas generadas, esto es,

$$\begin{aligned}\hat{I}_2 &= \frac{1}{J}\sum_{j=1}^{J}\hat{R}_{j2} \\ &= \frac{1}{8}\left(\hat{R}_{A2}+\hat{R}_{R2}+\cdots+\hat{R}_{O2}\right) \\ &= \frac{1}{8}\left(\frac{\hat{r}_{A2}}{\hat{r}_{A0}}+\frac{\hat{r}_{R2}}{\hat{r}_{R0}}+\cdots+\frac{\hat{r}_{O2}}{\hat{r}_{O0}}\right) \\ &= \frac{1}{8}\left(\frac{\hat{r}'_{A2}+\hat{r}'_{R2}+\cdots+\hat{r}'_{O2}}{\hat{r}'_{GT2}}\right)\end{aligned} \tag{3.22}$$

donde el subíndice 2 representa el año que se compara, en este caso 2002, con el año base, 2000. En forma análoga, se obtiene el índice de Kish, para el año 2004, $\hat{I}_4$.

El cálculo de este índice se hará en forma ilustrativa ya que, como se mencionó anteriormente, el INPC tiene una estructura más compleja de cálculo. El punto importante de estimar el índice propuesto por Kish es obtener la estimación de su varianza, lo cual resulta complejo debido a que se genera directamente de cocientes de estimadores de razón que provienen de muestras aleatorias diferentes (años diferentes) y para los cuales los diseños muestrales pueden variar en los distintos años de estudio considerados. Cabe mencionar que el índice mostrado en (3.22) no es un índice ponderado, esto es, a todos los términos que intervienen en el promedio les da el mismo peso, por lo que supone que todas las relativas tienen la misma aportación de información al promedio final.

3.2.1. Preparación de las Bases de Datos

Se hizo una revisión previa de la información de las bases de datos de la ENIGH, básicamente en los tamaños de muestra de cada año, para así determinar a qué nivel sería posible desagregar la información y adecuarla a los cálculos a realizar en la estimación de la varianza, de tal manera que los resultados presentados no perdieran la representatividad de la estimación debido a un tamaño de muestra muy pequeño. También se revisó que la información contenida en las bases de datos fuera suficiente para el cálculo del estimador Horvitz-Thompson (estimación π) de totales y de la estimación de sus varianzas. La información contenida en las ENIGH, se observó era suficiente para el caso de la estimación π de los totales, esto es, se cuenta con los factores de expansión y el valor de las variables de las cuales se desea obtener un estimador Horvitz-Thompson del total. Estas variables corresponden a las categorías generadas, mencionadas anteriormente. Sin embargo, se encontró que la información contenida en estos archivos no era suficiente para el caso del cálculo de estimación de varianzas, ya que INEGI no presenta las Unidades Primarias de Muestreo (UPM) y tampoco los estratos en las bases de datos de la ENIGH, necesarias en las expresiones presentadas anteriormente.

Varianza propuesta por Kish

El método propuesto por Kish, para el cálculo de varianza, requiere que la información esté agrupada en estratos, véanse ecuaciones (3.17), (3.18) y (3.7). Las bases de datos de la ENIGH no presentaban las unidades primarias de muestreo en ninguno de los tres años, tampoco se mencionaba claramente en el documento metodológico cuáles fueron los estratos para el levantamiento de la muestra por lo que se decidió posestratificar las bases de datos. Las variables consideradas para tal propósito fueron la entidad federativa (con 32 estratos) y el tipo de localidad (rural y urbano), haciendo un total de 64 estratos. Esta posestratificación se realizó en los archivos de los 3 años del gasto en hogares a nivel nacional debido a que las expresiones obtenidas así lo requerían, y porque mezclaban 2 años diferentes (muestras) dentro de un mismo estrato. Para el caso de las unidades primarias de muestreo, se decidió considerar hipotéticamente a los municipios existentes en cada uno de los estratos para la definición del diseño.

Una vez que se determinaron los estratos y se aplicaron a las bases de datos, se procedió a usar las expresiones dadas en las ecuaciones (3.12), (3.13) y (3.14) para estimar la varianza de una relativa. Cabe mencionar que el total de unidades de muestreo en la población para cada estrato, N_h, en cada uno de los distintos años (2000, 2002 y 2004) eran desconocidos, por lo que se decidió tomar la estimación de cada uno de ellos usando los correspondientes factores de expansión. Esto es $\hat{N}_{0h} = \sum_{s_{0h}} \frac{1}{\pi_k}$, $\hat{N}_{2h} = \sum_{s_{2h}} \frac{1}{\pi_k}$ y $\hat{N}_{4h} = \sum_{s_{4h}} \frac{1}{\pi_k}$ para $h = 1, \ldots, 64$, donde $\hat{N}_{0h}$, $\hat{N}_{2h}$ y $\hat{N}_{4h}$ son la estimaciones del total de unidades últimas de muestreo en el estrato h

para los años 2000, 2002 y 2004.

De igual forma, para obtener el estimador de la media muestral en la expresión (3.11) se requería conocer el total poblacional. Un estimador alterno a la media muestral es $\tilde{y}_{s_h} = \frac{1}{\hat{N}_h}\sum_{s_h}\frac{y_k}{\pi_k}$, véase Särndal (1992), el cual se usó una vez estimados todos los totales poblacionales por estrato y año, por lo que el estimador de la covarianza queda de la siguiente manera

$$\widehat{C}\left(\hat{r}_{0j},\hat{r}_{1k}\right)=\sum_{h=1}^{H}\frac{1}{\hat{t}_{x_{0j}}\hat{t}_{x_{1k}}}\left(\frac{\hat{N}_{0jh}\hat{N}_{1kh}}{\sqrt{n_{0jh}\left(n_{0jh}-1\right)}\sqrt{n_{1kh}\left(n_{1kh}-1\right)}}\right)\sum_{l=1}^{n_{0jh}}\sum_{i=1}^{n_{1kh}}d_{0jhl}d_{1khi}$$

donde

$$\begin{aligned}
d_{0jhl} &= \left(y_{0jhl}-\tilde{y}_{0jh}\right)-\hat{r}_{0j}\left(x_{0jhl}-\tilde{x}_{0jh}\right)\\
d_{1khi} &= \left(y_{1khi}-\tilde{y}_{1kh}\right)-\hat{r}_{1k}\left(x_{1khi}-\tilde{x}_{1kh}\right)\\
\tilde{y}_{1kh} &= \frac{1}{\hat{N}_{1kh}}\sum_{i=1}^{n_{kh}}\frac{y_{1khi}}{\pi_i}
\end{aligned}$$

y en forma equivalente para $\tilde{x}_{1kh}$, $\tilde{y}_{0jh}$ y $\tilde{x}_{0jh}$.

Después de haber adecuado las bases de datos, y haber definido los detalles de las variables de cálculo y del diseño, se procedió a programar en el paquete estadístico **R** parte por parte cada uno de los términos que intervienen en la varianza del índice de Kish, dados en (3.18). Más adelante se presentarán algunos de los resultados obtenidos.

Varianza Jackknife para el índice de Kish

Para aplicar la técnica jackknife en la estimación de la varianza del índice de Kish también se tuvieron que adecuar las bases de datos a las necesidades requeridas.

La técnica jackknife contempla eliminar un elemento a la vez y obtener el estimador tantas veces como elementos se tengan. Los elementos considerados aquí, o Unidades Primarias de Muestreo, fueron los municipios por entidad, por lo que para estimar la varianza jackknife, se definió primero el número de municipios por entidad federativa y después el número total de municipios en muestra a nivel nacional. Se tomó la entidad como parte del indicador de la UPM, debido a que la clave de los municipios en las bases de datos en ocasiones coincidía cuando se tomaban dos entidades federativas diferentes.

Una vez generadas las UPM, con clave de identificación única, se calcularon los índices propuestos por Kish considerando lo siguiente. El índice es un promedio en el cual cada sumando es el cociente de dos estimadores de razón, cada estimador de razón proviene de un año (muestra) diferente, por lo que para aplicar jackknife se tuvieron que generar dos vectores del índice de Kish, esto es, se calcularon $\hat{I}_{(01)}$, $\hat{I}_{(02)}, \ldots, \hat{I}_{(0n_0)}, \hat{I}_{(21)}, \hat{I}_{(22)}, \ldots, \hat{I}_{(2n_2)}$ donde cada término tiene la siguiente forma

$$\hat{I}_{(2j)} = \frac{1}{L}\sum_{k=1}^{L} \hat{R}_k^{(2j)} = \frac{1}{L}\sum_{k=1}^{L} \frac{\hat{r}_{2l}^{(j)}}{\hat{r}_0}, \qquad j = 1, \ldots, n_2$$

$$\hat{I}_{(0j)} = \frac{1}{L}\sum_{k=1}^{L} \hat{R}_k^{(0j)} = \frac{1}{L}\sum_{k=1}^{L} \frac{\hat{r}_{2l}}{\hat{r}_0^{(j)}}, \qquad j = 1, \ldots, n_0$$

el subíndice indica a qué año se le afectó con el procedimiento jackknife. Esto es, para $\hat{I}_{(0j)}$ se fueron quitando las unidades primarias de muestreo sólo en el estimador de razón del denominador, que es el correspondiente al año base, 2000, teniéndose así n_0 estimadores jackknife del índice de Kish. Lo mismo se hace para el año de estudio, en este caso 2002, teniéndose así n_2 estimadores jackknife. De esta manera se generaron los vectores

$$\hat{I}_{(0)} = \begin{bmatrix} \hat{I}_{(01)} \\ \hat{I}_{(02)} \\ \vdots \\ \hat{I}_{(0n_0)} \end{bmatrix}, \qquad \hat{I}_{(2)} = \begin{bmatrix} \hat{I}_{(21)} \\ \hat{I}_{(22)} \\ \vdots \\ \hat{I}_{(2n_2)} \end{bmatrix}$$

los cuales tienen los términos de la estimación del índice de Kish, quitando una UPM a la vez. A partir de estos vectores, se genera uno sólo de dimensión $(n_0 + n_2) \times 1$, esto es, $\hat{I}_{JK} = \begin{bmatrix} \hat{I}_{(0)} \\ \hat{I}_{(2)} \end{bmatrix}$. Con este vector es posible obtener la estimación jackknife de la varianza del índice de Kish mediante la expresión

$$\widehat{V}_{JK}\left(\hat{I}_2\right) = \frac{1}{(n_0+n_2)(n_0+n_2-1)} \sum_{j=1}^{n_0+n_2} \left[\hat{I}_{JK}(j) - \left(\frac{1}{n_0+n_2}\sum_{j=1}^{n_0+n_2} \hat{I}_{JK}(j)\right)\right]^2,$$

y cuyos resultados se muestran más adelante. En forma análoga se obtienen los resultados para el estimador de la varianza jackknife del índice de Kish para el año 2004, $\hat{I}_4$.

Grupos Aleatorios Dependientes (*Dependent Random Groups*)

Otro método para la estimación de la varianza, mencionado a detalle en Särndal et al. (1992), es conocido como Grupos Aleatorios Dependientes. Para aplicar este método, se asume que se tiene una muestra grande obtenida del total de la población mediante un diseño muestral probabilístico. Después que se obtuvo esta muestra, se genera un mecanismo aleatorio para dividirla en un número de submuestras ajenas que serán, los grupos aleatorios. Estos no serán independientes, sin embargo, serán tratados como si lo fueran. Estos grupos resultan ser heterogéneos a su interior debido a la forma en que los elementos fueron seleccionados. A continuación se hace una breve descripción del procedimiento para obtener tal estimador de la varianza.

Sea s la muestra de tamaño n obtenida de la población U, que será la muestra completa. Ahora, s se divide en A grupos aleatorios ajenos, $s_1, \ldots, s_a, \ldots, s_A$, esto es,

$$s = \bigcup_{a=1}^{A} s_a$$

Sea s de tamaño fijo n, supóngase que los grupos son de igual tamaño, $m = n/A$. Supóngase que s se divide en forma aleatoria, de tal manera que cada grupo aleatorio tiene el mismo diseño muestral que la muestra padre. Se selecciona $s_i,\ i = 1, \ldots, A$, muestra de elementos sin reemplazo. Sean $\hat{\theta}_1, \ldots, \hat{\theta}_a, \ldots, \hat{\theta}_A$ estimadores de θ, donde $\hat{\theta}_a$ es el estimador en la muestra de s_a, entonces se obtiene

$$\begin{aligned}
\hat{\theta}_{DRG} &= \frac{1}{A} \sum_{a=1}^{A} \hat{\theta}_a \\
\widehat{V}_{DRG} &= \frac{1}{A\,(A-1)} \sum_{a=1}^{A} \left(\hat{\theta}_a - \hat{\theta}_{DRG}\right)^2
\end{aligned}$$

donde $\widehat{V}_{DRG}$ es el estimador de varianza obtenido por este método.

El inconveniente de aplicar este método al caso práctico del índice de Kish es que se tienen dos muestras provenientes de años diferentes. Además, el tamaño de muestra es diferente en cada año considerado, por lo que no fue posible generar directamente los grupos aleatorios en cada encuesta.

3.3. Resultados Numéricos Obtenidos y Conclusiones

En las tablas B.3, B.4 y B.5 del Anexo, se presentan los tamaños de muestra por entidad y tamaño de la localidad considerados en la ENIGH. En la gráfica 3.1, se presentan los tamaños de muestra de la ENIGH para los años 2000, 2002 y 2004 por entidad federativa. Como se puede apreciar, la mayoría de las entidades del país, en el año 2000, tienen un tamaño de muestra menor a 500 unidades, sólo el estado de Veracruz presenta un tamaño de muestra mayor. Para el año 2002, 14 entidades tienen un tamaño de muestra menor a 500, 16 entidades entre 500 y mil unidades y sólo dos presentan un tamaño de muestra mayor de mil. En el año 2004, 13 entidades

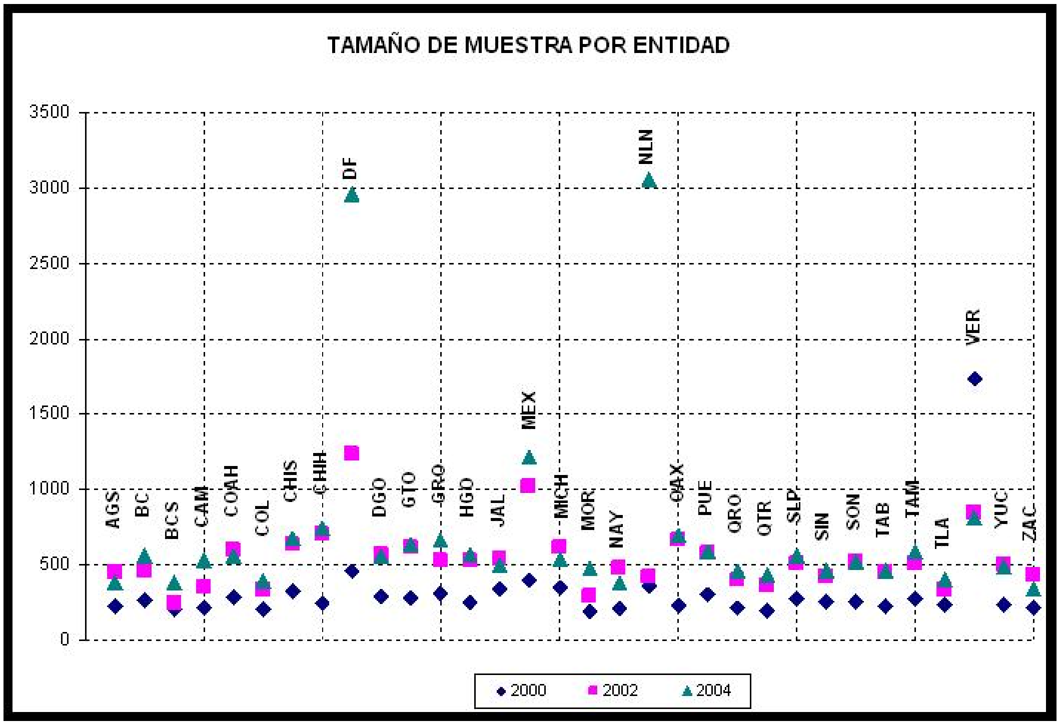

Figura 3.1: Tamaños de Muestra correspondiente a las ENIGH 2000, 2002 y 2004 por Entidad Federativa.

tienen un tamaño de muestra menor a 500, 16 entidades de 500 a mil unidades y el Estado de México, Distrito Federal y Nuevo León mayor a mil. Estas tres entidades presentan un tamaño de muestra considerable, por lo que, en caso de querer presentar la información con la variable de desagregación Entidad Federativa, ésta sólo sería representativa para las entidades federativas mencionadas, lo cual no lo hace comparable con los otros años de estudio. Además, en la Síntesis Metodológica de la ENIGH 2004, véase [25] se mencionan los niveles de desagregación en los cuales se pueden obtener estimaciones sin ningún problema de representatividad. Estos son a nivel nacional y urbano-rural. Por esta razón se decidió no hacer el desglose de la información considerando la Entidad Federativa.

En la tabla B.6 se presenta el tamaño de muestra por entidad, pero desglosado en Urbano-Rural. Esta desagregación es la que sirvió para los cálculos de la varianza del Indice de Kish. En la gráfica 3.2, se tienen los tamaños de muestra de la ENIGH correspondiente a los años 2000, 2002 y 2004, pero ahora considerando a la distribución urbano rural[3] del país y el nacional. En este caso, los tamaños de muestra para las localidades rurales son muy parecidos en los tres años, aunque cabe mencionar que este tamaño es pequeño con respecto a la muestra considerada en las localidades

[3]Definición considerada por el INEGI. Rural: localidades con menos de 2,500 habitantes; Urbana: localidades de 2,500 y más habitantes.

urbanas del país. Aún así, la muestra en comunidades rurales es lo suficientemente grande para la obtención de resultados confiables. En el caso de nivel urbano, los tamaños de muestra son mayores que para el rural, y debe observarse que entre años se presentan diferencias considerables en el tamaño de muestra. Esto no afecta en los resultados finales, pero se debe tener presente que un cambio en la estimación de varianzas se puede deber al tamaño de la muestra para cada año considerado. Finalmente, se decidió usar esta desagregación de la información debido a que los tamaños de muestra presentaban mejores condiciones en los grupos generados que en el caso de entidad federativa.

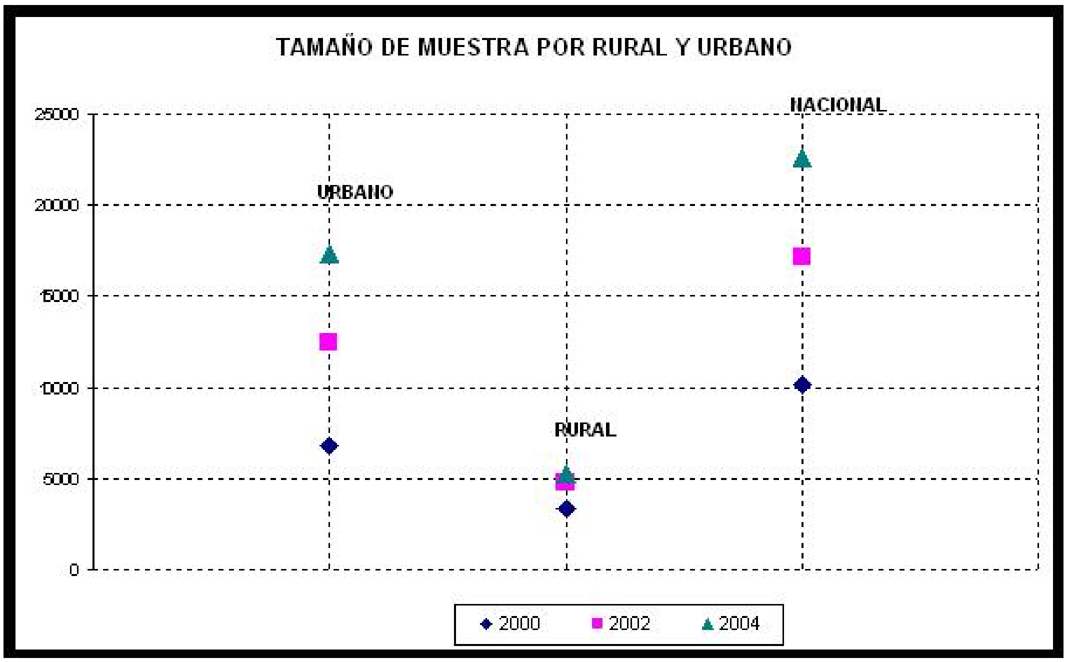

Figura 3.2: Tamaños de Muestra correspondiente a las ENIGH 2000, 2002 y 2004, considerando la distribución por Urbano y Rural y a nivel nacional.

Para el análisis de la información, se obtuvo primero la estimación del Gasto Mensual Promedio Total proveniente de la ENIGH en los años 2000, 2002 y 2004. Las variables referentes al gasto fueron agrupadas en cada una de las categorías anteriormente mencionadas, y con las cuales se trabajó para obtener los estimadores de razón por categoría en los años 2002 y 2004, tomando como base el 2000. Además de los estimadores de razón en los cuales el denominador es el Gasto Total, ecuación (3.19). De manera similar se obtiene la estimación de las relativas por categoría para los años 2002 y 2004, considerando como base el año 2000. Finalmente, se obtuvo la estimación del índice de Kish para los años 2002 y 2004. El diseño muestral considerado para el cálculo de las estimaciones fue un muestreo estratificado en el cual la entidad federativa y la población rural-urbana fueron las variables de estratificación. El diseño original de la ENIGH, para los años 2000, 2002 y 2004, es un polietápico, estratificado y por conglomerados, véase [23], [24] y [25]. El inconveniente fue que en la ENIGH 2004 no se conocían con exactitud las variables usadas para estratificar, debido a que no se mencionaba en la Síntesis Metodológica de la ENIGH 2004.

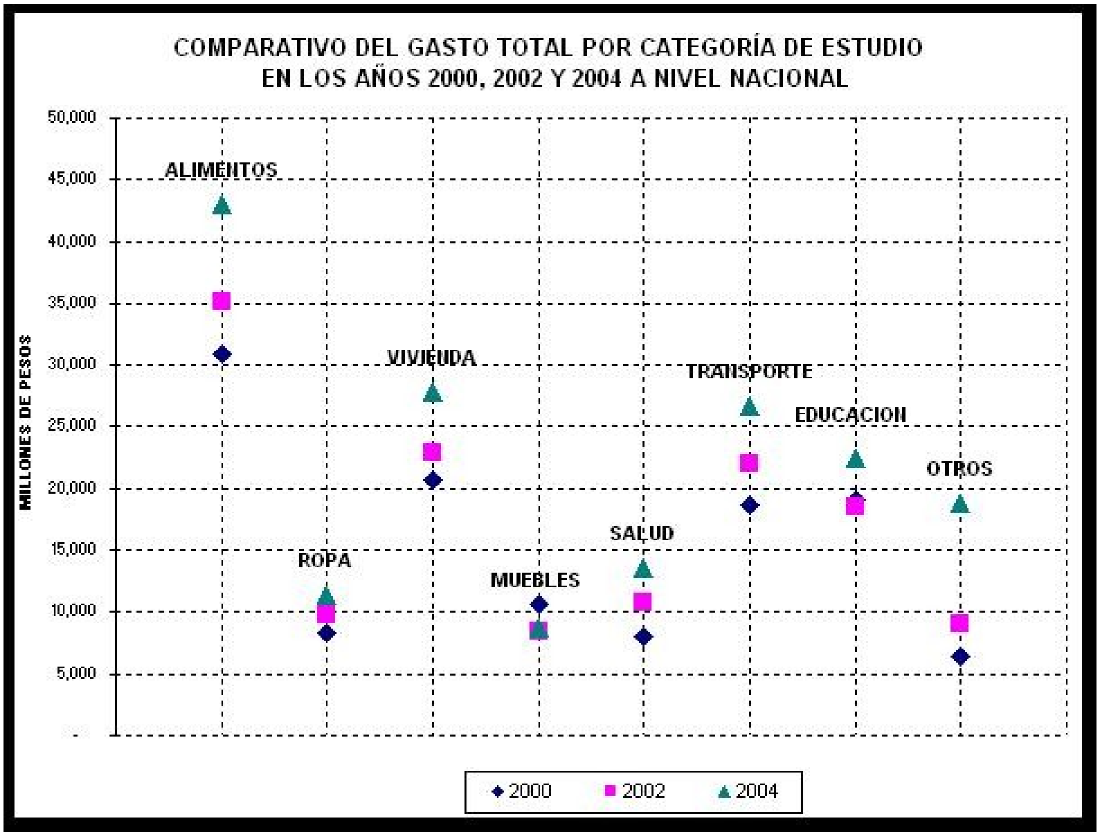

Figura 3.3: Gasto Total estimado en cada una de las categorías consideradas para los años de estudio 2000, 2002 y 2004 a nivel Nacional.

En la tabla 3.2 se muestra el gasto, anteriomente mencionado, para el nivel nacional y los tres años de estudio. Se aprecia que el Gasto Total estimado aumentó respecto al año base (2000) y al año anterior (2004 respecto a 2002), a pesar de esto, en algunas de las categorías consideradas el gasto correspondiente disminuyó respecto al año base. Estos cambios se aprecian mejor en la tabla 3.3, en la cual se presentan los estimadores de razón de los años 2000-2002 y 2000-2004 con su respectiva estimación de la desviación estándar y el intervalo de confianza del estimador de razón. La estimación de la varianza se obtuvo mediante dos métodos; el que presenta Kish en su artículo (1968) y usando la técnica jackknife al estimador del índice de Kish. De estos resultados, se aprecia que la estimación de la varianza por jackknife dio intervalos de confianza de rango mayor a los obtenidos por Kish. Esto se aprecia con mayor claridad en la tabla 3.6, donde se muestra la estimación del error estándar obtenida por ambos métodos, así como el cociente de ambas estimaciones, para el año 2002 y 2004 respecto al 2000. Se observa en las tablas 3.2 y 3.3 el cambio que hubo en el gasto mensual promedio total en hogares para cada una de las categorías mencionadas, siendo en Muebles y Educación donde el gasto disminuyó del 2000 al 2002 y sólo en Muebles para el año 2000 al 2004. En la gráfica 3.3 se muestra la estimación puntual del Gasto Total por categoría para los años mencionados.

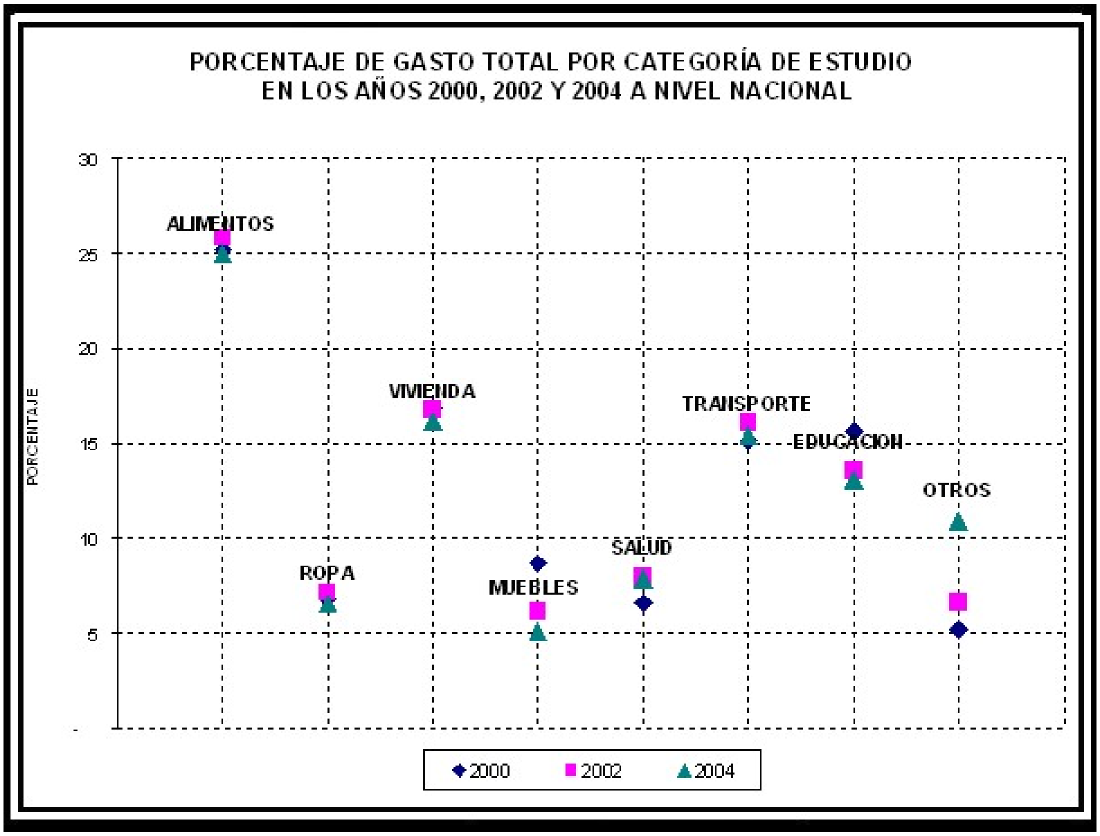

Figura 3.4: Porcentaje del Gasto Total estimado en cada una de las categorías consideradas para los años de estudio 2000, 2002 y 2004 a nivel Nacional.

En la tabla 3.4 se muestran los resultados de las estimaciones de las relativas, $\hat{R}_2$ y $\hat{R}_4$, con su respectiva estimación de la desviación estándar obtenida por el método presentado por Kish y la técnica jackknife, así como la estimación del índice de Kish. La comparación de estos resultados se muestra a detalle en la tabla 3.7, para los años 2002 y 2004. Para el caso del 2002, no es posible decir cuál método fue más efectivo debido a que en la categoría de **ROPA** el intervalo de confianza presenta mayor rango por el método de estimación de Kish, mientras que para el resto de las categorías la estimación del error estándar fue menor por el método de Kish. Para el 2004, la estimación del error estándar de las relativas fue mayor usando la técnica jackknife. En general, la técnica jackknife produjo estimaciones de varianza de mayor magnitud que las obtenidas por Kish. Al final de cada tabla, también se muestra la estimación del índice de Kish, dado en la ecuación (3.22). Para este caso, la estimación de la varianza obtenida con la técnica jackknife fue más chica en ambos años que la que propone Kish, pero no se puede concluir si es mejor estimación que la obtenida con el otro método, debido a que no se conoce la varianza verdadera con la cual sería posible compararla. Esto hace más difícil definir la eficacia de ambos métodos y saber cuál resulta ser mejor ya que se obtuvieron resultados similares para el caso de urbano y rural. Además, el no contar con la varianza original de cada relativa, y del índice, impide determinar con cuál método se obtuvieron resultados más precisos.

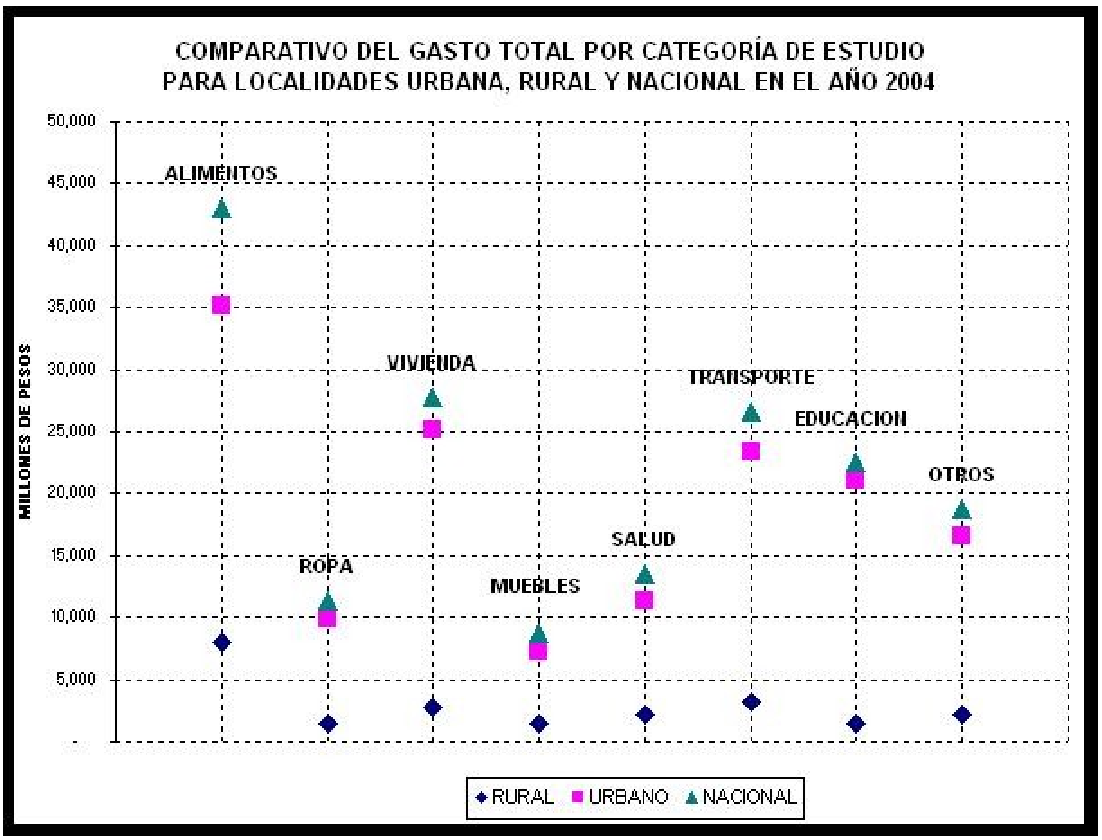

Figura 3.5: Gasto Total estimado en cada una de las categorías consideradas para localidades Rurales, Urbanas y a nivel Nacional en el año 2004.

En la tabla 3.5 se muestran los estimadores de razón por categoría. Estos estimadores están calculados con respecto al Gasto Promedio Mensual en hogares mexicanos en los años 2000, 2002 y 2004. A diferencia de los estimadores mostrados en la tabla 3.3, estos no consideran como punto de referencia el año 2000, por lo que no es posible definir cuál de ellos tuvo mayores cambios en el tiempo. Por ejemplo, en el año 2004 hubo un mayor gasto en la categoría **OTROS**, haciendo que disminuyera en **MUEBLES**, **VIVIENDA**, **EDUCACIÓN** y **ROPA**, respecto al gasto total en hogares. Sin embargo, si se considera el cambio respecto al año 2000, el decremento sólo se dio en **MUEBLES**. Es por eso que la forma de interpretación de estas tablas debe hacerse con cuidado para no obtener conclusiones erradas. La estimación de la varianza para estos estimadores de razón se hizo mediante el método descrito en Särndal et al. (1992).

Como se mencionó anteriormente, el desglose de la información no es representativa a nivel entidad federativa, por lo que sólo se muestran resultados a nivel Urbano-Rural. En la tabla C.1 se presenta la estimación del Gasto Promedio Total a nivel Urbano, agregada para las 8 categorías de estudio. En la tabla C.2 se presentan los estimadores de razón de los años 2002 y 2004 a nivel Urbano y para cada una de las categorías. En esta tabla se aprecia el mismo comportamiento que para el nivel Nacional, en las categorías **MUEBLES** y **EDUCACIÓN** hubo un decremento del gasto total del año 2000 al 2002, mientras que para el año 2000 al 2004 sólo fue en **MUEBLES**. En cuanto a la estimación de la varianza, ocurre lo mismo que a nivel nacional. En la tabla C.5 se muestra la comparación de la estimación de ambos métodos.

En la tabla C.7 se muestran los mismos estimadores de totales pero para la población Rural. La tabla C.8 muestra los estimadores de razón correspondientes a la tabla anterior. Aquí se aprecia que sólo hubo un decremento en el gasto total en la categoría de **MUEBLES** para ambos años considerados, 2002 y 2004, respecto al 2000, mientras que en la categoría **OTROS** hubo un incremento muy significativo. En este caso, debe tomarse en cuenta que el tamaño de muestra considerado para las poblaciones rurales, el cual se vio en la gráfica 3.2, era muy pequeño respecto al nacional. Además, debe considerarse que en esta categoría en las poblaciones rurales el gasto es muy pequeño por el tipo de productos y servicios que comprende, en comparación con las poblaciones urbanas. La gráfica 3.5 muestra este comportamiento por categoría y población urbana-rural. Por estas razones, se obtienen estimaciones de las varianzas para los estimadores de razón muy grandes, comparadas con las de poblaciones urbanas. La estimación de la varianza de ambos métodos, para la población rural, es muy similar a los resultados obtenidos a nivel nacional y urbano. Con la técnica jackknife se obtuvieron estimaciones de varianzas para las relativas más grandes que con el método propuesto por Kish. Sólo en el caso del 2002 la estimación jackknife para el índice de Kish fue mayor, opuesto a los resultados obtenidos a nivel nacional y urbano. En las tablas C.11 y C.12 se muestra la comparación de ambos métodos de estimación de varianza para la población Rural.

En la tabla C.13 se presenta la estimación del Gasto Total para la Cd. de México y Zona Metropolitana del Estado de México[4] desglosada por las categorías antes

[4]El término *zona metropolitana*, CONAPO (2003), refiere a una forma particular de urbanización en donde la expansión de la ciudad hacia su periferia tiende a rebasar los límites territoriales

vistas. Los municipios y delegaciones que se consideran en esta zona se presentan en la tabla B.1. Cabe mencionar que el método propuesto por Kish no se puso aplicar a esta zona debido a que, como se mencionó, se considera la estratificación de la muestra para los años que intervienen en el cálculo. La estratificación realizada fue sobre las entidades federativas y tipo de localidad urbano rural, quedando partida esta zona con los estratos e incompleta para el caso del Estado de México y la población rural en el Distrito Federal para uno de los años considerados, haciendo el proceso incompatible con la información. En la tabla C.14 están los estimadores de razón para el Gasto Total, mencionado anteriormente. La estimación de la varianza se hizo sólo con la técnica jackknife.

En la tabla C.15 se muestran los estimadores de razón del Gasto Promedio Mensual para las distintas categorías respecto al gasto total en el hogar, la estimación de la varianza se hizo mediante el método descrito en Särndal et al. (1992). En esta tabla se aprecia un comportamiento similar que para los dominios Rural, Urbano y Nacional, el Gasto Total en hogares creció en los años más recientes para la mayoría de las categorías, mientras que **MUEBLES** bajó en ambos años de estudio.

En general, los resultados obtenidos para la relativas presentaron una estimación de varianza menor con el método propuesto por Kish, caso contrario ocurrió con la estimación de la varianza del índice de Kish, ya que la técnica jackknife resultó dar varianzas menores. Si se contara con la varianza original de las relativas y del índice, se sabría claramente cuál de los dos métodos fue mejor, por lo que sólo se pudo determinar con cuál de los métodos se obtenían varianzas menores.

Una observación importante a los resultados obtenidos es que a éstos no se les corrigió el factor de la inflación generado en los años de estudio, por lo que al corregir la información los resultados finales obtenidos podrían variar, en forma similar a los cambios inflacionarios.

En las secciones §3.3.1 §C.1.1 §C.1.2 y §C.1.3 se presentan los resultados de la estimación de varianza obtenidos mediante el método propuesto por Kish en su artículo (1968), además de la estimación de las varianzas para las relativas y el índice de Kish, obtenidas mediante la técnica jackknife.

En los resultados del índice de Kish presentados en las secciones mencionadas arriba, se consideró el factor que aporta la covarianza entre las encuestas de los distintos años, esto es, los cálculos se hicieron bajo el supuesto de que las ENIGH no eran independientes. Se realizó el mismo ejercicio suponiendo que eran indepen-

---

de la unidad político-administrativa que originalmente la contenía, incorporando como parte de sí misma o de su área de influencia directa a unidades político-administrativas vecinas, con las que constituye un ámbito territorial altamente integrado física y funcionalmente. Para más información véase [3].

dientes, con covarianza cero, obteniéndose resultados similares. En la tabla 3.1 se presenta la estimación del índice de Kish a nivel Nacional y Urbano Rural para los años 2002 y 2004 con sus respectivas estimaciones del error estándar, considerando que las encuestas ENIGH eran dependientes con el año 2000 y considerando independencia entre ellas. La diferencia entre los estimadores de la desviación estándar se observó, básicamente, en el tercer decimal, por lo que los intervalos de confianza obtenidos no tuvieron grandes diferencias, presentándose, finalmente, en este trabajo sólo los resultados en los que se considera dependencia de las encuestas ENIGH, esto es, covarianza distinta de cero.

Las gráficas de los resultados obtenidos en este capítulo se anexan en el apéndice §C.

Tabla 3.1: Estimación de la desviación estándar del índice de Kish a nivel Nacional, Urbano y Rural para los años 2002 y 2004, considerando encuestas ENIGH dependientes y posteriormente independientes.

2002			
	INDICE KISH	**CONSIDERANDO COVARIANZA**	**SIN CONSIDERAR COVARIANZA**
URBANO	1.0156	0.0186	0.0192
RURAL	1.1994	0.1521	0.1525
NACIONAL	1.0228	0.0234	0.0238

2004			
	INDICE KISH	**CONSIDERANDO COVARIANZA**	**SIN CONSIDERAR COVARIANZA**
URBANO	1.0766	0.0235	0.0236
RURAL	1.1919	0.0418	0.0386
NACIONAL	1.0807	0.0233	0.0234

3.3.1. Estimaciones a Nivel Nacional

Tabla 3.2: Estimación del Gasto Mensual Promedio en Hogares Mexicanos a nivel **Nacional** por categoría, en los años 2000, 2002 y 2004. Información obtenida de la ENIGH 2000, 2002 y 2004. Las categorías son las consideradas en el INPC.

CATEGORIA	**GASTO TOTAL DE LOS AÑOS**		
	2000	**2002**	**2004**
ALIMENTOS	30,924,035,829	35,181,337,410	43,060,007,789
ROPA	8,339,131,611	9,701,773,164	11,351,496,881
VIVIENDA	20,653,247,229	22,871,833,989	27,825,017,376
MUEBLES	10,640,878,376	8,392,457,369	8,786,196,656
SALUD	8,055,897,422	10,782,021,816	13,568,509,044
TRANSPORTE	18,637,322,462	21,950,015,330	26,663,118,166
EDUCACION	19,110,217,524	18,450,688,662	22,472,605,714
OTROS	6,407,056,141	9,031,653,781	18,771,944,708
GASTO TOTAL	122,767,786,593	136,361,781,521	172,498,896,333

Tabla 3.3: Estimación de Razón del Gasto Mensual Promedio a nivel **Nacional** por categoría, para los años 2002, $\hat{r}_2'\left(=\frac{\hat{t}_{A02}}{\hat{t}_{A00}}\right)$, y 2004, $\hat{r}_4'\left(=\frac{\hat{t}_{A04}}{\hat{t}_{A00}}\right)$, así como estimación de su desviación estándar, $se\left(\hat{r}_2'\right)$ y $se\left(\hat{r}_4'\right)$, y límites de confianza de la estimación al 95 %, considerando el método propuesto por Kish (1968) y empleando la técnica jackknife, tomando como base de comparación el año 2000.

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{r}_2'$	$se\left(\hat{r}_2'\right)$	**INF**	**SUP**	$se\left(\hat{r}_2'\right)$	**INF**	**SUP**
ALIMENTOS	113.77	10.90	92.41	135.13	14.33	85.68	141.85
ROPA	116.34	13.20	90.46	142.22	16.72	83.57	149.11
VIVIENDA	110.74	13.63	84.02	137.46	17.65	76.15	145.34
MUEBLES	78.87	8.12	62.95	94.78	10.35	58.58	99.16
SALUD	133.84	13.58	107.21	160.47	18.17	98.22	169.46
TRANSPORTE	117.77	18.04	82.41	153.14	21.54	75.55	159.99
EDUCACION	96.55	21.87	53.68	139.41	27.41	42.82	150.28
OTROS	140.96	23.23	95.44	186.49	27.61	86.84	195.08
GASTO TOTAL	111.07	13.37	84.86	137.29	16.90	77.96	144.19

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{r}_4'$	$se\left(\hat{r}_4'\right)$	**INF**	**SUP**	$se\left(\hat{r}_4'\right)$	**INF**	**SUP**
ALIMENTOS	139.24	12.98	113.81	164.68	17.21	105.52	172.97
ROPA	136.12	15.07	106.59	165.66	19.36	98.17	174.07
VIVIENDA	134.72	16.45	102.47	166.97	20.74	94.07	175.38
MUEBLES	82.57	8.21	66.47	98.67	10.55	61.89	103.25
SALUD	168.43	18.01	133.13	203.73	22.85	123.65	213.21
TRANSPORTE	143.06	20.47	102.95	183.18	24.95	94.17	191.96
EDUCACION	117.59	27.21	64.26	170.93	32.82	53.27	181.92
OTROS	292.99	37.78	218.95	367.03	46.99	200.89	385.09
GASTO TOTAL	140.51	16.61	107.96	173.06	20.76	99.83	181.19

Tabla 3.4: Estimación de las Relativas del Gasto Mensual Promedio a nivel **Nacional** para cada una de las categorías consideradas en el INPC, y estimación del índice de Kish para los años 2002, $\hat{R}_2 \left(= \frac{\hat{r}_2}{\hat{r}_0}\right)$, y 2004, $\hat{R}_4 \left(= \frac{\hat{r}_4}{\hat{r}_0}\right)$, así como estimación de su desviación estándar, $se\left(\hat{R}_2\right)$ y $se\left(\hat{R}_4\right)$, y límites de confianza al 95 % de la estimación, considerando el método propuesto por Kish (1968) y empleando la técnica jackknife, tomando como base de comparación el año 2000.

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{R}_2$	$se\left(\hat{R}_2\right)$	**INF**	**SUP**	$se\left(\hat{R}_2\right)$	**INF**	**SUP**
ALIMENTOS	1.0243	0.0547	0.9170	1.1315	0.0607	0.9052	1.1433
ROPA	1.0474	0.0540	0.9415	1.1533	0.0531	0.9433	1.1516
VIVIENDA	0.9970	0.0398	0.9191	1.0749	0.0428	0.9132	1.0809
MUEBLES	0.7101	0.0464	0.6191	0.8010	0.0490	0.6141	0.8060
SALUD	1.2050	0.0626	1.0822	1.3277	0.0660	1.0757	1.3343
TRANSPORTE	1.0603	0.0585	0.9456	1.1751	0.0632	0.9364	1.1843
EDUCACION	0.8692	0.1210	0.6320	1.1065	0.1471	0.5809	1.1576
OTROS	1.2691	0.1389	0.9969	1.5414	0.1437	0.9874	1.5508
INDICE KISH	**1.0228**	**0.0265**	**0.9708**	**1.0748**	**0.0132**	**0.9969**	**1.0487**

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{R}_4$	$se\left(\hat{R}_4\right)$	**INF**	**SUP**	$se\left(\hat{R}_4\right)$	**INF**	**SUP**
ALIMENTOS	0.9910	0.0532	0.8867	0.9910	0.0541	0.8849	1.0953
ROPA	0.9688	0.0450	0.8806	0.9688	0.0458	0.8791	1.0570
VIVIENDA	0.9588	0.0384	0.8835	0.9588	0.0402	0.8800	1.0342
MUEBLES	0.5877	0.0361	0.5169	0.5877	0.0380	0.5131	0.6585
SALUD	1.1987	0.0588	1.0835	1.1987	0.0617	1.0778	1.3139
TRANSPORTE	1.0182	0.0465	0.9270	1.0182	0.0510	0.9182	1.1094
EDUCACION	0.8369	0.1198	0.6021	0.8369	0.1407	0.5611	1.0718
OTROS	2.0852	0.1518	1.7876	2.0852	0.1569	1.7777	2.3828
INDICE KISH	**1.0807**	**0.0269**	**1.0280**	**1.0807**	**0.0178**	**1.0458**	**1.1333**

Tabla 3.5: Estimación de Razón del Gasto Mensual Promedio a nivel **Nacional** para cada una de las categorías respecto al Gasto Total, para los años 2000, $\hat{r}_0$, 2002, $\hat{r}_2$, y 2004, $\hat{r}_4$. Estimación de su desviación estándar, $se\,(\hat{r}_0)$, $se\,(\hat{r}_2)$ y $se\,(\hat{r}_4)$, y límites de confianza al 95 % de la estimación.

CATEGORIA	$\hat{r}_0$	$se\,(\hat{r}_0)$	**INF**	**SUP**
ALIMENTOS	25.19	1.23	22.78	27.60
ROPA	6.79	0.30	6.21	7.37
VIVIENDA	16.82	0.60	15.64	18.01
MUEBLES	8.67	0.50	7.69	9.65
SALUD	6.56	0.31	5.96	7.16
TRANSPORTE	15.18	0.66	13.88	16.48
EDUCACION	15.57	2.12	11.41	19.72
OTROS	5.22	0.36	4.52	5.92

CATEGORIA	$\hat{r}_2$	$se\,(\hat{r}_2)$	**INF**	**SUP**
ALIMENTOS	25.80	0.78	24.28	27.32
ROPA	7.11	0.16	6.80	7.43
VIVIENDA	16.77	0.33	16.12	17.43
MUEBLES	6.15	0.18	5.79	6.52
SALUD	7.91	0.20	7.51	8.31
TRANSPORTE	16.10	0.59	14.93	17.26
EDUCACION	13.53	0.50	12.54	14.52
OTROS	6.62	0.57	5.50	7.74

CATEGORIA	$\hat{r}_4$	$se\,(\hat{r}_4)$	**INF**	**SUP**
ALIMENTOS	24.96	0.51	23.95	25.97
ROPA	6.58	0.10	6.39	6.77
VIVIENDA	16.13	0.28	15.58	16.68
MUEBLES	5.09	0.10	4.90	5.29
SALUD	7.87	0.15	7.57	8.17
TRANSPORTE	15.46	0.30	14.86	16.05
EDUCACION	13.03	0.44	12.17	13.88
OTROS	10.88	0.29	10.32	11.45

Tabla 3.6: Estimación del error estándar, $se\left(\hat{r}_2'\right)$ y $se\left(\hat{r}_4'\right)$, para los estimadores de razón, $\hat{r}_2'\left(=\frac{\hat{t}_{A02}}{\hat{t}_{A00}}\right)$ y $\hat{r}_4'\left(=\frac{\hat{t}_{A04}}{\hat{t}_{A00}}\right)$, a nivel **Nacional** obtenidos mediante el método propuesto por: **KISH.** (1968) y **JK.** la técnica jackknife aplicada al estimador de razón. Además, cociente de la estimación del error estándar de los métodos mencionados para los años 2002 y 2004, respectivamente.

	2002/2000			**2004/2000**		
CATEGORIA	**KISH**	**JK**	**JK/KISH**	**KISH**	**JK**	**JK/KISH**
ALIMENTOS	10.90	14.33	1.31	12.98	17.21	1.33
ROPA	13.20	16.72	1.27	15.07	19.36	1.28
VIVIENDA	13.63	17.65	1.29	16.45	20.74	1.26
MUEBLES	8.12	10.35	1.27	8.21	10.55	1.28
SALUD	13.58	18.17	1.34	18.01	22.85	1.27
TRANSPORTE	18.04	21.54	1.19	20.47	24.95	1.22
EDUCACION	21.87	27.41	1.25	27.21	32.82	1.21
OTROS	23.23	27.61	1.19	37.78	46.99	1.24
GASTO TOTAL	13.37	16.90	1.26	16.61	20.76	1.25

Tabla 3.7: Estimación del error estándar, $se\left(\hat{R}_2\right)$ y $se\left(\hat{R}_4\right)$, para los estimadores de las relativas, $\hat{R}_2\left(=\frac{\hat{r}_2}{\hat{r}_0}\right)$, y $\hat{R}_4\left(=\frac{\hat{r}_4}{\hat{r}_0}\right)$, y del índice de Kish, $\hat{I}_2\left(=\frac{1}{J}\sum_{j=1}^{J}\hat{R}_{j2}\right)$ e $\hat{I}_4\left(=\frac{1}{J}\sum_{j=1}^{J}\hat{R}_{j4}\right)$, a nivel **Nacional** obtenidos mediante el método propuesto por: **KISH.** (1968) y **JK.** la técnica jackknife aplicada al estimador de las relativas y al índice de Kish. Además, cociente de la estimación del error estándar de los métodos mencionados para los años 2002 y 2004, respectivamente.

	2002/2000			**2004/2000**		
CATEGORIA	**KISH**	**JK**	**JK/KISH**	**KISH**	**JK**	**JK/KISH**
ALIMENTOS	0.0547	0.0607	1.11	0.0532	0.0541	1.02
ROPA	0.0540	0.0531	0.98	0.0450	0.0458	1.02
VIVIENDA	0.0398	0.0428	1.08	0.0384	0.0402	1.05
MUEBLES	0.0464	0.0490	1.06	0.0361	0.0380	1.05
SALUD	0.0626	0.0660	1.05	0.0588	0.0617	1.05
TRANSPORTE	0.0585	0.0632	1.08	0.0465	0.0510	1.10
EDUCACION	0.1210	0.1471	1.22	0.1198	0.1407	1.17
OTROS	0.1389	0.1437	1.03	0.1518	0.1569	1.03
INDICE KISH	**0.0265**	**0.0132**	**0.50**	**0.0269**	**0.0178**	**0.66**

Capítulo 4

Conclusiones

La estimación de varianzas para el caso en que se tienen estimadores no lineales se hace complejo debido a que en ocasiones no se cuenta con una expresión explícita de la misma que permita su cálculo.

La finalidad de este trabajo fue revisar la estimación de algunos parámetros de interés que en la práctica no siempre son presentados con una estimación de su error estándar, o varianza, por lo que fue importante ilustrar la metodología que podrían seguir. Tales parámetros son: la Mediana de una población, el índice de Gini y el índice de Kish. En cada uno de los casos, se ilustró la forma de cómo obtener un estimador puntual a partir de información muestral de distintas fuentes, tales como el XII Censo General de Población y Vivienda del 2000 y la Encuesta Nacional de Ingresos y Gastos en los Hogares, en sus ediciones del 2000, 2002 y 2004.

Esta información también sirvió para obtener la estimación de la varianza de los parámetros mencionados. Para la Mediana, la estimación se hizo utilizando el paquete estadístico **SUDAAN**, mediante las fórmulas de los límites de confianza presentadas por Binder (1991) y Francisco y Fuller (1991) y que inicialmente fueron propuestas por Woodruff (1952).

La estimación del índice de Gini y su varianza, se programaron inicialmente en el paquete estadístico **S**-plus, para la estimación a nivel entidad federativa; posteriormente, se migró la información al paquete estadístico **R** debido a que **S**-plus no soportó la base de datos de la muestra del censo para obtener la estimación del índice de Gini a nivel nacional. En este caso, se obtuvo la estimación del índice de Gini por dos métodos, uno propuesto por Ogwang (2000) que obtiene el estimador a partir de mínimos cuadrados ponderados y estima la varianza del índice de Gini usando la técnica jackknife. Ogwang no hace uso de ningún diseño muestral en particular al presentar las expresiones de estimación del índice. El otro método que se usó es el propuesto por Sandström et al (1988), el cual considera los factores de expansión del diseño muestral en la estimación del índice de Gini, además, propone una estimación de su varianza. Para este caso, se obtuvo además la estimación del índice de Gini, y de su varianza, suponiendo un diseño muestral aleatorio simple, esto es, considerando los factores de expansión constantes. Los resultados obtenidos por este método se compararon con los del método de Ogwang, teniendo estimación de varianza más pequeña cuando se utilizó el método de Sandström aplicado a un m.a.s., aunque cabe mencionar que las diferencias no fueron muy considerables. La razón mayor de los errores estándar estuvo en 1.12, de Ogwang respecto a Sanström, para el estado de Oaxaca, mientras que la menor fue de 1.03 en Baja California, y para el nacional fue de 1.05.

Para el mismo método, Sandström et al., se obtuvo la estimación de la varianza mediante la técnica jackknife. En este caso, no fue posible distinguir un método que, en general, diera estimación de varianza más pequeña que el otro, ya que en algunas entidades federativas la varianza obtenida mediante la técnica jackknife fue menor que la de Sandström y viceversa. Para el caso de la estimación de varianza del índice de Gini nacional, no hubo diferencias, ya que ambos métodos dieron la misma estimación.

La estimación del índice de Kish y su varianza se programaron también en el paquete estadístico **R**. La dificultad de este método radicó en que se manejó información de encuestas diferentes, haciendo más difícil la manipulación de la información. Se obtuvo la estimación del índice de Kish a nivel nacional y población rural y urbana, así como la estimación de su varianza considerando el caso en que las encuestas no eran independientes y el caso en que lo eran, teniéndose resultados muy parecidos, por lo que los cálculos que se realizaron consideraron el factor de la covarianza en la estimación de la varianza, esto es, se supuso que las encuestas no eran independientes. Además de la estimación de varianza propuesta por Kish, se calculó la estimación de varianza usando la técnica jackknife. Los resultados obtenidos usando esta técnica, en general dieron estimaciones de varianza más grande que las obtenidas por el método de Kish, tanto para los estimadores de razón como para las relativas. Sólo en el caso del índice de Kish, la estimación de su varianza por jackknife resultó ser menor que la obtenida por el método propuesto. Las estimaciones se obtuvieron a nivel nacional y por tipo de población rural y urbana, obteniéndose resultados similares en los tres casos.

La decisión de usar el software mencionado en la estimación de estos parámetros fue porque algunos de ellos son de acceso gratuito en la red, tal es el caso de **R**, o porque son versiones estudiantiles que sirvieron para los propósitos alcanzados en este trabajo. Además de los mencionados aquí, existen otros como **SAS**, **STATA**, **SPSS** o **WesVar**, el inconveniente de estos paquetes, por ejemplo **SAS**, es que son muy caros en el mercado y no cuentan con los detalles de programación que fueron necesarios para obtener la estimación de los parámetros aquí mencionados. Tampoco cuentan con una ventana de comandos en la cual se puedan programar rutinas definidas por el usuario y que realicen cálculos específicos. La mayoría de ellos cuenta con funciones o expresiones predeterminadas y que el usuario no puede modificar o adecuar al problema en sí.

Ahora, si un usuario desea obtener la estimación del índice de Gini y de su varianza, por ejemplo, se le recomendaría que usara el método propuesto por Sandström considerando diseño muestral, esto debido a que si se considera un diseño muestral aleatorio simple los resultados puntuales también cambian, siendo más parecidos a los que se obtuvieron por Ogwang. La técnica jackknife es recomendable en ciertos casos, y se recomienda no usar la técnica clásica a un muestreo estratificado por ejemplo, ya que los estimadores cambian dependiendo del diseño muestral. Ahora que si se desea obtener el resultado mediante la técnica jackknife, se debe ser cuidadoso con el diseño muestral al momento de obtener los estimadores.

De igual forma para la estimación del índice de Kish y de su varianza, es recomendable usar las expresiones dadas y que ya han sido previamente verificadas por Kish, en vez de usar la técnica jackknife, sobretodo porque al mezclar dos encuestas diferentes los problemas de estimación crecieron.

Si el propósito es sistematizar tales métodos de estimación de varianza, se recomienda seguir los pasos metodológicos de las expresiones que ya han sido verificadas por los autores mencionados y tomar la técnica jackknife como una alternativa de estimación, sin ser esta técnica la principal para resolver el problema. Así, si se sistematiza el método general de Sandström, se sabe, de este trabajo particular, que si se aplica un muestreo aleatorio simple, entonces los resultados serán similares a los obtenidos por Ogwang, y que si se obtiene por jackknife, entonces habrá estimaciones de varianza que sobreestimen su valor y otras que la subestimen. También para el índice de Kish será recomendable contar con la sistematización de la estimación, sobretodo por el problema que genera la manipulación de la información, y que una vez sistematizado, será fácil obtener tales resultados.

Una recomendación adicional que se hace es tener especial atención en el proceso de sistematización en el caso de que un tercero sea el encargado del mismo, ya que los métodos de estimación de varianza aquí presentados tienen expresiones complicadas que requieren de un entendimiento más profundo para facilitar el proceso y no tener errores durante el proceso.

Apéndice A

Uso de software

A.1. Paquete estadístico SUDAAN

Para la estimación de la Mediana del Ingreso Promedio en hogares, se utilizó el paquete estadístico **SUDAAN 7.5.6**. Este paquete resultó de gran ayuda porque en forma automática obtiene también la estimación del error estándar de la estimación de la Mediana. Además, la sintaxis que emplea no es muy compleja. Este paquete tiene la ventaja de leer Bases de Datos con formato en código **ASCII** (y algunas modificaciones a los mismos), **SPSS** (versión 5.x y posteriores), **SAS**. También tiene la cualidad de exportar Bases de Datos de código **ASCII** a tipo **SAS**. Véase [12].

La sintaxis que tienen los procesos es de muy fácil manejo ya que no necesita de muchas opciones. A continuación se listan algunas de las opciones que forman parte de la sintaxis usada para el cálculo de estimación de la mediana y de su varianza. Posteriormente, se presenta un ejemplo que sirvió como base para obtener la tabla 1.1.

- **PROC**. Esta palabra indica que se realizará un procedimiento, el cual se especifica después de la declaración de esta palabra.
- **DESCRIPT**. Esta se define inmediatamente después de la palabra **PROC**. Con este procedimiento se producen información descriptiva para el análisis de variables continuas y categóricas. Esta información incluye la estimación de medias, medianas y otros cuantiles y porcentajes, además de la estimación de su desviación estándar de encuestas sobre muestras y otros estudios que envuelven datos correlacionados.
- **DATA**. Con esta opción se especifica el archivo del cual se extraerá la información. La ruta del archivo debe estar entre comillas (“”) y cada directorio se debe especificar con doble contradiagonal (\\).
- **FILETYPE**. Aquí se especifica el tipo de archivo externo que se leerá, **SPSS**, **SAS**, etc.

- **DESIGN**. Con esta opción se especifica el tipo de diseño muestral con el que se recolectó la información y mediante el cual se realizará la estimación de los parámetros. Cuenta con varias alternativas, las cuales son `SRS` para un muestreo aleatorio simple, `WR` para un muestreo con reemplazo, `WOR` para un muestreo sin reemplazo, `UNEQWOR` para un muestreo sin reemplazo con iguales probabilidades de selección en la primera etapa. También se pueden especificar aquí los métodos de estimación de la varianza en caso de que se aplique un método de remuestreo, las opciones son `JACKKNIFE` y `BRR`.

- **NEST**. Con la declaración de esta opción se listan las variables que identificarán o etiquetarán los niveles o etapas usadas en el diseño muestral, estas variables deben listarse en el mismo orden que la muestra fue seleccionada. Una de las opciones de `NEST` es `/MISSUNIT`, con la cual se especifica que cuando sólo se encuentra una unidad muestral dentro de una etapa determinada, la contribución de la varianza de esta unidad se estimará usando la diferencia en ese valor de la unidad y el valor de la media general de la población.

- **WEIGHT**. Con esta opción se especifica la variable de los factores de expansión del diseño muestral. Dependiendo el diseño muestral, especificado en **DESIGN**, esta opción puede o no aparecer.

- **PERCETILE**. En esta opción se especifica el parámetro a estimar, en el caso de este trabajo la mediana. Se antepone una diagonal a la definición del parámetro.

- **VAR**. Los parámetros de esta opción debe ser la lista de variables de las cuales se obtendrá la estimación especificada en el punto anterior.

- **SUBGROUPS**. Con esta opción se declara(n) la(s) variable(s) con el nombre de las categorías en las cuales se realizará la estimación de la información.

- **LEVELS**. Con esta opción se especifica el número de categorías que tiene cada una de las variables declaradas en `SUBGROUPS`.

Es importante poner siempre **;** al final de casa opción que se lista en la sintaxis, de lo contrario el proceso no se ejecutará.

A continuación se presenta un ejemplo de la sintaxis usada para la estimación de la Mediana del Ingreso Promedio Mensual.

Con el primer ejemplo, se obtiene la estimación puntual y un intervalo de confianza al 95 % de la Mediana del Ingreso Mensual Promedio por Hogares (IMPH), nacional y para las 32 entidades federativas, considerando un diseño muestral complejo. La estimación de la varianza usada por este paquete es por aproximación de series de Taylor. Véase tabla 1.1.

```
PROC DESCRIPT DATA=‘‘D:\\MUESTREO\\VHONALXG.SAV’’ FILETYPE=SPSS DESIGN=WR;
NEST ESTRATOX UPMX /MISSUNIT;
WEIGHT FACTORX;
PERCENTILE /MEDIAN;
VAR INGTOHOX;
SUBGROUPS ENTX;
LEVELS 32;
```

Donde `VHONALXG.SAV` es la base de datos en formato **SPSS**, las variables de estratificación son `ESTRATOX` y `UPMX`, los factores de expansión es `FACTORX`, la variable de la que se calculará la estimación de la mediana es INGTOHOX, la cual estará desagregada en los niveles que tenga de la variable `ENTX`, que se indica con `LEVELS` que son 32.

En el segundo ejemplo se obtiene la estimación puntual y un intervalo de confianza al 95 % de la Mediana del Ingreso Mensual Promedio por Hogares (IMPH), nacional y para las 32 entidades federativas considerando un diseño muestral Aleatorio Simple (MAS). El estimador de la varianza usado por este paquete es por aproximación de series de Taylor. Nótese en este ejemplo que muchas de las declaraciones hechas en el ejemplo anterior ya no aparecen, esto debido a que el tipo de diseño muestral especificado en este segundo ejemplo no las requiere.

```
PROC DESCRIPT DATA=‘‘D:\\MUESTREO\\VHONALX.SAV’’ FILETYPE=SPSS DESIGN=SRS;
PERCENTILE /MEDIAN;
VAR INGTOHOX;
SUBGROUPS ENTX;
LEVELS 32;
```

A.2. Paquete estadístico R

Para el propósito del cálculo de estimación de varianzas de los parámetros presentados en la sección §3.3, se utilizó el módulo estadístico **survey**, V.3.3-1, del paquete **R 2.2.0** (diseñado por Thomas Lumley, Profesor Asociado de la Universidad de Washington, Seattle), que está enfocado básicamente al análisis y estimación de parámetros de encuestas que tienen un diseño muestral complejo (diferente del aleatorio simple). Este procedimiento de estimación incluye resúmenes estadísticos, modelos lineales generalizados, estimación general máximo pseudo-verosímil para muestreos estratificados multietápicos, muestreo de conglomerados, muestreo de encuestas con distintos pesos; además de que la estimación de la varianza la hace aplicando el método de linealización de series de Taylor.

Se optó por el uso de esta herramienta estadística, la cual resulta ser bastante robusta para los propósitos del presente trabajo, debido a que es un software de uso libre que se puede encontrar en el web, y que constantemente se actualiza o tiene nuevos módulos de cálculo, no sólo en el ámbito estadístico sino en distintas áreas del conocimiento.

La metodología de estimación con que cuentan algunas de las opciones de este módulo está basada principalmente en Särndal et al. (1992), entre otros.

La estimación de la varianza de los índices mediante el método propuesto por Kish (1968), véase sección §3.1, se programó casi en su totalidad, debido a que las opciones presentadas en este módulo estadístico no eran suficientes para obtener tales resultados.

Los programas generados fueron varios, además de extensos, debido a que se tenían que calcular las estimaciones primero a nivel Nacional, después a nivel Urbano-Rural y, finalmente, se hizo el ejercicio para la Zona Metropolitana del D.F. y Estado de México. La estimación de la varianza se obtuvo usando dos métodos, la aproximación dada por Kish y usando la técnica Jackknife.

A continuación se presentan algunos de los comandos de **survey** usados para la estimación de la varianza de los estimadores de razón, de totales y del índice de Kish.

- **svytotal**. Esta función ayudó a obtener la estimación π de los totales para cada una de las variables (categorías) generadas, para cada uno de los años de estudio. La notación usada es de la siguiente manera, donde `ALIMENTOS`, `ROPA`, `VIVIENDA`,..., `GAS_TOT` son las variables de estudio y `D.00` es el diseño muestral que se define previamente a la estimación de los totales. En este caso, el vector de resultados se almacena en la variable `T.00`.

```
T.00 <- svytotal(~ALIMENTOS + ROPA + VIVIENDA + MUEBLES + SALUD
     + TRANSPORTE + EDUCACION + OTROS + GAS_TOT,D.00)
```

- **vcov**. Con esta función se obtiene la matriz de varianzas-covarianzas de las variables especificadas. Para el uso de esta función es necesario usar previamente **svytotal** o **svymean** (éste último obtiene un estimador de medias de las variables indicadas). Al objeto resultante en la opción anterior se le aplica esta función, obteniendo así la matriz de datos de varianzas y covarianzas. Nota: Por sí sola la función **vcov** no trabaja. La notación usada para tal opción es la siguiente, donde la variable `T.00`, obtenida arriba, tiene la estimación de los totales de las variables mencionadas, y de las cuales se obtendrá su matriz de varianzas-covarianzas, y se almacenará en el objeto `VC.00`.

```
VC.00 <- vcov(T.00)
```

- **svyratio**. Esta función sirvió para obtener los estimadores de razón, y su respectiva varianza, de cada una de las variables (categorías) consideradas respecto al Gasto Mensual Promedio Total. Nota: Estos estimadores son los correspondientes a las tablas 3.5, C.4, C.10 y C.15, en los cuales la estimación de los totales tanto del numerador como del denominador pertenecen al mismo año. La estimación de razón en las tablas 3.3, C.2, C.8, C.14 se obtuvo programando los resultados paso a paso, debido a que numerador y denominador provienen de encuestas y años diferentes.

```
V.00 <- svyratio(~ALIMENTOS + ROPA + VIVIENDA + MUEBLES + SALUD
     + TRANSPORTE + EDUCACION + OTROS,~GAS_TOT, D.00)
```

El único inconveniente encontrado en el uso de esta última opción es que no se puede obtener la covarianza entre dos estimadores de razón, es decir, no se puede usar la opción **vcov** para obtener la matriz de varianzas-covarianzas, la cual es necesaria para el caso de la estimación de varianzas de los índices de Kish.

A continuación se presenta una parte del sintaxis que sirvió para obtener algunos de los resultados presentados en la sección §3.3. La sintaxis de **R** es muy parecida a la usada por el paquete estadístico **S**-Plus.

En el sintaxis descrito, primero se definen las librerías que ayudarán en los procesos de estimación y en el manejo de las bases de datos. En caso de no estar instaladas en **R**, se pueden obtener en forma gratuita de la página web citada en [17]. Se recomienda bajar el archivo en formato `*.zip` e instalarlo desde el menú de **R**.

Las tres librerías usadas, en este caso, son: **survey**, que ayuda en la estimación de parámetros de muestras complejas; **foreign**, que ayuda a importar datos externos que no son hechos en **R** y exportar datos generados en **R**; **session**, que guarda en un archivo ajeno de la memoria de **R** las variables y tablas generadas en la sesión actual. Es de suma importancia que la opción **session** se use para guardar los resultados en un archivo independiente a **R**, debido a que si no se hace esto, la memoria de **R** poco a poco se satura haciendo el cálculo de las opciones programadas más lento, e incluso que no se puedan terminar los procesos por falta de capacidad de memoria. En el ejemplo presentado, se pone la opción **restore.session** y la ruta en la que se encuentra el archivo de la sesión generado, en este caso, `KISH.RData`. Para generar este archivo, se usa la opción **save.session** y la ruta donde será almacenado en disco duro. Al final de la sintaxis mostrada se pone el ejemplo de cómo debe usarse.

Para leer la base de datos se usó la opción **read.table** con los respectivos parámetros de lectura. Estos pueden cambiar dependiendo el formato de la base de datos. Puede estar separado por comas, o sin encabezados. En este caso, en el primer renglón se tenía el nombre de la variable y cada variable estaba separada por tabuladores.

Nota: Para la lectura de archivos, se cuenta también con las opciones **read.csv** que permite la lectura de archivos de texto plano separadas las variables con comas y la opción **read.spss** que permite leer archivos en forma directa con formato de **SPSS**, sin necesidad de exportarlos con otro formato. Esto es de gran ayuda cuando se tienen bases de datos con este tipo de formatos. Sólo bastaría revisar la sintaxis de cada una de las opciones.

Como se tenía la información de las tres ENIGH, 2000, 2002 y 2004 usadas en un mismo archivo, lo primero fue separarlos con un nombre diferente. La opción presentada en la sintaxis, después de la lectura de la base de datos y asignada a `INPC`, filtra primero la información para el año indicado en la base `INPC` y posteriormente la asigna a la base `INPCXX`, donde `XX` es el año seleccionado.

Una vez separada la información en cada ENIGH, se define el diseño muestral usando la opción **svydesign**, función propia del módulo **survey**, y se asigna a una variable, en este caso `D.00`. Algunos de los parámetros básicos que debe llevar esta opción son **id** (en el ejemplo `LOCALIDAD`), que es el identificador en caso de que existan conglomerados en los registros; **variables** (`estrato`, `region`, `ALIMENTOS`,..., `GAS_TOT`, `ANIO`), que es donde se definen todas las variables que intervendrán en el diseño muestral; **weight** (`factor`), que son los ponderadores (o factores de expansión) del diseño muestral; **data** (`INPC00`), que es el nombre de la base de datos definida. Adicional a estos parámetros, la definición del diseño muestral puede llevar variables de estratificación (**strata**), probabilidades de los conglomerados (**probs**), que pueden ser en una o varias etapas, dependiendo si es un vector o una matriz de datos, y factores de corrección por finitud (**fpc**), de igual forma un vector o matriz de datos.

Una vez definido el diseño muestral, es posible utilizar las opciones de estimación de **survey**.

En el ejemplo, en la variable `V.00` se almacenan los resultados generados por la opción **svyratio**. Aquí se calcula el estimador de razón para todas las variables definidas con el diseño muestral `D.00`, que es el correspondiente al año 2000. Las variables consideradas en el numerador del estimador de razón se definen primero, después del símbolo $\sim$ y separadas con un signo $+$, después se especifican las variables del denominador con las cuales se cruzarán las previamente definidas para el numerador, de igual forma después del símbolo $\sim$, en caso de haber varias variables en el denominador se usa el signo $+$ para especificarlas.

Con la siguiente opción del ejemplo, **svytotal**, se obtiene la estimación de los totales para las variables especificadas. Aquí la sintaxis es más sencilla, después del símbolo $\sim$ se escriben las variables a estimar su total, separadas con un signo $+$. En forma automática **R** identifica cada una de las variables consideradas. A continuación, se pone el nombre del diseño muestral, `D.00`, que en este caso es el correspondiente al año 2000. Los resultados se almacenan en la tabla `T.00`.

Una vez obtenidas las estimaciones de los totales, se obtiene la matriz de covarianzas de las variables especificadas. Abajo de esta opción, se presenta una parte del programa que calcula la estimación de la matriz de varianzas y covarianzas para estimadores de razón. Esta ecuación fue obtenida de Kish (1979) y que no es parte del módulo **survey**. Fue programada usando los resultados obtenidos con el módulo **survey**, mencionados anteriormente, y usando la forma de multiplicación matricial con la que cuenta **R**, que está definida por $A\% * \%B$, donde A y B son matrices de dimensión $(n_1 \times n_2)$ y $(n_2 \times n_3)$. Posterior a esto, se hicieron varios cálculos, que por la magnitud del programa no se presentan en el documento. Los resultados obtenidos se almacenaron en la variable `I.KISH` para posteriormente ser guardados en disco, usando la opción **write.table** en el archivo con nombre `I-KISH-GRAL.DAT`, y separado por comas.

```
#____________________________________________
#
#  LIBRERIAS NECESARIAS
#____________________________________________
#
 library(survey)
 library(foreign)
 library(session)

  restore.session('D:\\MUESTREO\\ENIGH\\VAR-IKISH-GRAL.RData')


#____________________________________________
#
#  ABRE LA BASE DE DATOS GENERAL
#____________________________________________
#
   INPC <- read.table('D:\\MUESTREO\\ENIGH\\INPC.DAT',
     header=TRUE,sep = '''', dec=''.'', nrows=-1)

#   SEPARA LOS DATOS POR EL ANIO DE LA ENIGH

 INPC00<-INPC[INPC$ANIO==2000,]
```

```
 INPC02<-INPC[INPC$ANIO==2002,]
 INPC04<-INPC[INPC$ANIO==2004,]

# SE DEFINE EL DISEGNO MUESTRAL, SOLO PARA 2000.

    D.00 <- svydesign(id=~LOCALIDAD, strata =~ENTIDAD + RURURB,
      variables =~estrato + region
       + ENTIDAD + TPOB + RURURB
       + ALIMENTOS + ROPA + VIVIENDA + MUEBLES
       + SALUD + TRANSPORTE + EDUCACION + OTROS
       + GAS_TOT + ANIO,
      weight =~factor,
      nest = TRUE,
      data = INPC00)
...

# ---> VARIANZAS DE ESTIMADORES DE RAZON PARA CADA CATEGORIA.

 V.00 <- svyratio(~ALIMENTOS+ ROPA + VIVIENDA + MUEBLES + SALUD
   + TRANSPORTE + EDUCACION + OTROS,~GAS_TOT, D.00)
...

# ---> ESTIMADORES DEL TOTAL POR CATEGORIA, ANIO 2000.

T.00 <- svytotal(~ALIMENTOS+ ROPA + VIVIENDA + MUEBLES + SALUD
   + TRANSPORTE + EDUCACION + OTROS + GAS_TOT, D.00)
...

# MATRIZ DE VAR-COV DE T.00.

 VC.00 <- vcov( T.00 )

# --->  FORMULA OBTENIDA DE KISH, PARA COVARIANZAS DE DOS ESTIMADORES DE RAZON.

 VC.rj.rk.00 <- ((VC.00[1:8,1:8] + (V.00$ratio %*% t(V.00$ratio))*VC.00[9,9]
   - ( matrix(rep(V.00$ratio * VC.00[1:8,9],8),8,8))
   - t( matrix(rep(V.00$ratio * VC.00[1:8,9],8),8,8)))/T.00[9]^2)
...

# CON LOS RESULTADOS OBTENIDOS SE GENERA VARIABLE I.KISH Y SE
# GUARDA EN DISCO CON LA SIGUIENTE OPCION.
#
# (NO SE MUESTRA EXPRESION DEBIDO A LA COMPLEJIDAD DE LA MISMA)
```

```
  write.table(I.KISH.sinC,file='D:\\MUESTREO\\ENIGH\\
  I-KISH-GRAL-sinC.DAT',sep=",")

# SE GUARDAN LAS VARIABLES GENERADAS EN LA SESION POR SEPARADO.

save.session('D:\\MUESTREO\\ENIGH\\VAR-IKISH-GRAL.RData')
```

Apéndice B

Información de la Encuesta ENIGH

Tabla B.1: Delegaciones del DF y Municipios o Localidades del Estado de México consideradas como la Zona Metropolitana de la Ciudad de México, según la definición para el cálculo del INPC.

DELEGACIONES DEL DF		**MUNICIPIOS O LOCALIDADES DEL ESTADO DE MÉXICO**			
1	Azcapotzalco	1	Acolman	17	La Paz
2	Coyoacán	2	Atizapán de Zaragoza	18	Tecamac
3	Cuajimalpa	3	Coacalco	19	Teoloyucan
4	Gustavo A. Madero	4	Cuautitlán	20	Tepotzotlán
5	Iztacalco	5	Chalco	21	Tlalnepantla
6	Iztapalapa	6	Chicoloapan	22	Tultepec
7	Magdalena Contreras	7	Chimalhuacán	23	Tultitlán
8	Milpa Alta	8	Ecatepec	24	Cuautitlán Izcalli
9	Alvaro Obregón	9	Huixquilucan	25	Valle de Chalco
10	Tlahuac	10	Ixtapaluca		
11	Tlalpan	11	Jaltenco		
12	Xochimilco	12	Melchor Ocampo		
13	Benito Juárez	13	Naucalpan de Juárez		
14	Cuauhtémoc	14	Nezahualcoyotl		
15	Miguel Hidalgo	15	Nextlalpan		
16	Venustiano Carranza	16	Nicolás Romero		

Tabla B.2: Número de Localidades obtenidas en la muestra por Entidad Federativa para la ENIGH de los años 2000, 2002 y 2004.

ENTIDAD	**2000**	**2002**	**2004**
AGS	8	9	7
BC	4	5	4
BCS	3	4	4
CAM	6	7	9
COAH	11	15	16
COL	6	6	9
CHIS	10	22	19
CHIH	7	19	15
DF	15	16	16
DGO	11	14	13
GTO	12	18	14
GRO	12	17	21
HGO	10	18	18
JAL	18	21	19
MEX	23	40	39
MICH	15	21	18
MOR	10	8	14
NAY	8	9	11
NLN	18	16	36
OAX	15	35	30
PUE	14	29	22
QRO	7	9	9
QTR	6	6	7
SLP	12	14	12
SIN	7	9	10
SON	7	12	11
TAB	8	10	11
TAM	13	13	13
TLA	17	12	21
VER	46	29	30
YUC	12	17	14
ZAC	9	15	11
TOTAL	**380**	**495**	**503**

Tabla B.3: Tamaño de Muestra por Entidad Federativa y Tamaño de la Localidad para la ENIGH 2000.

ENTIDAD	**>100,000**	**15,000-99,999**	**2,500-14,999**	**<2500**	**TOTAL**
AGS	82	31	31	82	226
BC	170	20	38	36	264
BCS	88	17	51	53	209
CAM	78	19	33	88	218
COAH	83	70	0	136	289
COL	22	64	35	82	203
CHIS	148	37	17	123	325
CHIH	107	55	0	85	247
DF	403	10	0	43	456
DGO	92	18	59	125	294
GTO	97	73	3	115	288
GRO	117	49	31	114	311
HGO	83	18	32	121	254
JAL	100	51	27	167	345
MEX	206	52	38	106	402
MICH	114	49	62	128	353
MOR	56	41	17	83	197
NAY	78	30	20	85	213
NLN	133	20	75	139	367
OAX	64	41	53	80	238
PUE	97	76	51	84	308
QRO	45	41	60	68	214
QTR	71	18	32	73	194
SLP	91	17	37	130	275
SIN	123	20	19	90	252
SON	116	33	19	83	251
TAB	68	16	42	98	224
TAM	131	26	37	85	279
TLA	0	50	91	91	232
VER	919	247	223	346	1735
YUC	45	50	59	80	234
ZAC	49	59	16	87	211
TOTAL	**4076**	**1418**	**1308**	**3306**	**10108**

Tabla B.4: Tamaño de Muestra por Entidad Federativa y Tamaño de la Localidad para la ENIGH 2002.

ENTIDAD	**>100,000**	**15,000-99,999**	**2,500-14,999**	**<2500**	**TOTAL**
AGS	157	99	38	157	451
BC	360	52	28	17	457
BCS	181	34	0	32	247
CAM	184	0	72	100	356
COAH	290	204	0	105	599
COL	131	159	18	29	337
CHIS	181	87	84	288	640
CHIH	364	54	34	251	703
DF	984	45	120	83	1232
DGO	232	64	53	223	572
GTO	320	148	0	149	617
GRO	181	71	94	188	534
HGO	87	112	71	256	526
JAL	191	86	85	179	541
MEX	559	114	163	185	1021
MICH	186	173	7	247	613
MOR	164	41	46	40	291
NAY	225	19	53	183	480
NLN	199	71	56	94	420
OAX	100	167	134	268	669
PUE	105	96	131	248	580
QRO	101	75	100	122	398
QTR	247	20	33	65	365
SLP	216	36	39	215	506
SIN	187	41	69	124	421
SON	236	127	40	121	524
TAB	216	51	66	122	455
TAM	216	197	16	82	511
TLA	0	203	48	81	332
VER	285	244	82	229	840
YUC	118	101	140	142	501
ZAC	81	139	71	137	428
TOTAL	**7284**	**3130**	**1991**	**4762**	**17167**

Tabla B.5: Tamaño de Muestra por Entidad Federativa y Tamaño de la Localidad para la ENIGH 2004.

ENTIDAD	**>100,000**	**15,000-99,999**	**2,500-14,999**	**<2500**	**TOTAL**
AGS	187	90	0	106	383
BC	309	106	37	103	555
BCS	159	108	0	113	380
CAM	215	79	71	162	527
COAH	354	108	15	81	558
COL	123	169	0	97	389
CHIS	199	83	112	279	673
CHIH	385	91	56	211	743
DF	2810	28	36	90	2964
DGO	298	26	85	145	554
GTO	286	186	0	164	636
GRO	214	100	51	302	667
HGO	137	183	19	230	569
JAL	236	82	35	146	499
MEX	787	153	119	155	1214
MICH	223	121	57	139	540
MOR	189	121	20	150	480
NAY	192	66	0	129	387
NLN	2182	256	211	406	3055
OAX	140	147	112	295	694
PUE	154	178	72	183	587
QRO	161	110	53	141	465
QTR	200	75	35	125	435
SLP	244	110	0	203	557
SIN	210	70	35	143	458
SON	235	141	0	143	519
TAB	182	112	43	123	460
TAM	333	128	0	132	593
TLA	0	187	95	122	404
VER	325	255	36	195	811
YUC	170	151	16	157	494
ZAC	77	121	40	107	345
TOTAL	**11916**	**3941**	**1461**	**5277**	**22595**

Tabla B.6: Tamaño de Muestra por Entidad Federativa y Población Rural- Urbana, según la definición dada por INEGI.

ENTIDAD	**2000**		**2002**		**2004**	
	URB	**RUR**	**URB**	**RUR**	**URB**	**RUR**
AGS	144	82	294	157	277	106
BC	228	36	440	17	452	103
BCS	156	53	215	32	267	113
CAM	130	88	256	100	365	162
COAH	153	136	494	105	477	81
COL	121	82	308	29	292	97
CHIS	202	123	352	288	394	279
CHIH	162	85	452	251	532	211
DF	413	43	1149	83	2874	90
DGO	169	125	349	223	409	145
GTO	173	115	468	149	472	164
GRO	197	114	346	188	365	302
HGO	133	121	270	256	339	230
JAL	178	167	362	179	353	146
MEX	296	106	836	185	1059	155
MICH	225	128	366	247	401	139
MOR	114	83	251	40	330	150
NAY	128	85	297	183	258	129
NLN	228	139	326	94	2649	406
OAX	158	80	401	268	399	295
PUE	224	84	332	248	404	183
QRO	146	68	276	122	324	141
QTR	121	73	300	65	310	125
SLP	145	130	291	215	354	203
SIN	162	90	297	124	315	143
SON	168	83	403	121	376	143
TAB	126	98	333	122	337	123
TAM	194	85	429	82	461	132
TLA	141	91	251	81	282	122
VER	1389	346	611	229	616	195
YUC	154	80	359	142	337	157
ZAC	124	87	291	137	238	107
TOTAL	**6802**	**3306**	**12405**	**4762**	**17318**	**5277**

Apéndice C

Resultados Numéricos y Gráficos del Indice de Kish

C.1. Resultados Numéricos

C.1.1. Estimaciones a Nivel Urbano

Tabla C.1: Estimación del Gasto Mensual Promedio en Hogares Mexicanos a nivel **Urbano** por categoría, en los años 2000, 2002 y 2004. Información obtenida de la ENIGH 2000, 2002 y 2004. Las categorías son las consideradas en el INPC.

CATEGORIA	**GASTO TOTAL DE LOS AÑOS**		
	2000	**2002**	**2004**
ALIMENTOS	26,061,603,263	29,461,208,010	35,136,363,581
ROPA	7,388,148,533	8,381,096,542	9,887,792,378
VIVIENDA	19,195,767,780	20,689,493,313	25,103,938,141
MUEBLES	9,134,324,306	7,163,836,338	7,336,530,808
SALUD	6,967,614,407	9,117,314,824	11,399,123,975
TRANSPORTE	17,022,929,778	19,080,100,257	23,445,243,197
EDUCACION	18,295,488,124	17,007,354,429	21,009,593,799
OTROS	5,986,388,832	7,527,102,633	16,606,388,057
GASTO TOTAL	110,052,265,023	118,427,506,346	149,924,973,937

Tabla C.2: Estimación de Razón del Gasto Mensual Promedio a nivel **Urbano** por categoría, para los años 2002, $\hat{r}'_2 \left(= \frac{\hat{t}_{A02}}{\hat{t}_{A00}}\right)$, y 2004, $\hat{r}'_4 \left(= \frac{\hat{t}_{A04}}{\hat{t}_{A00}}\right)$, así como estimación de su desviación estándar, $se\left(\hat{r}'_2\right)$ y $se\left(\hat{r}'_4\right)$, y límites de confianza de la estimación al 95 %, considerando el método propuesto por Kish (1968) y empleando la técnica jackknife, tomando como base de comparación el año 2000.

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{r}'_2$	$se\left(\hat{r}'_2\right)$	**INF**	**SUP**	$se\left(\hat{r}'_2\right)$	**INF**	**SUP**
ALIMENTOS	113.04	13.04	87.48	138.60	16.36	80.98	145.11
ROPA	113.44	14.71	84.61	142.27	18.15	77.87	149.01
VIVIENDA	107.78	14.14	80.07	135.49	18.15	72.21	143.36
MUEBLES	78.43	9.43	59.94	96.92	11.57	55.75	101.10
SALUD	130.85	15.50	100.47	161.24	20.03	91.59	170.11
TRANSPORTE	112.08	18.55	75.72	148.45	21.97	69.01	155.16
EDUCACION	92.96	21.99	49.86	136.05	27.69	38.70	147.22
OTROS	125.74	17.91	90.63	160.84	22.74	81.16	170.32
GASTO TOTAL	107.61	14.41	79.37	135.85	17.89	72.55	142.67

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{r}'_4$	$se\left(\hat{r}'_4\right)$	**INF**	**SUP**	$se\left(\hat{r}'_4\right)$	**INF**	**SUP**
ALIMENTOS	134.82	15.28	104.88	164.76	19.23	97.14	172.50
ROPA	133.83	16.98	100.56	167.11	21.21	92.26	175.41
VIVIENDA	130.78	17.37	96.73	164.83	21.56	88.52	173.04
MUEBLES	80.32	9.40	61.89	98.74	11.57	57.64	103.00
SALUD	163.60	20.45	123.52	203.68	25.01	114.59	212.61
TRANSPORTE	137.73	21.75	95.10	180.36	26.15	86.47	188.99
EDUCACION	114.83	27.87	60.21	169.46	33.69	48.80	180.87
OTROS	277.40	39.25	200.47	354.33	48.07	183.18	371.62
GASTO TOTAL	136.23	18.23	100.51	171.95	22.24	92.63	179.83

Tabla C.3: Estimación de las Relativas del Gasto Mensual Promedio a nivel **Urbano** para cada una de las categorías consideradas en el INPC, y estimación del índice de Kish para los años 2002, $\hat{R}_2\left(=\frac{\hat{r}_2}{\hat{r}_0}\right)$, y 2004, $\hat{R}_4\left(=\frac{\hat{r}_4}{\hat{r}_0}\right)$, así como estimación de su desviación estándar, $se\left(\hat{R}_2\right)$ y $se\left(\hat{R}_4\right)$, y límites de confianza al 95 % de la estimación, considerando el método propuesto por Kish (1968) y empleando la técnica jackknife, tomando como base de comparación el año 2000.

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{R}_2$	$se\left(\hat{R}_2\right)$	**INF**	**SUP**	$se\left(\hat{R}_2\right)$	**INF**	**SUP**
ALIMENTOS	1.0505	0.0572	0.9384	1.1626	0.0638	0.9254	1.1756
ROPA	1.0542	0.0602	0.9363	1.1721	0.0588	0.9389	1.1694
VIVIENDA	1.0016	0.0434	0.9166	1.0866	0.0458	0.9119	1.0913
MUEBLES	0.7288	0.0519	0.6270	0.8306	0.0545	0.6220	0.8357
SALUD	1.2160	0.0634	1.0918	1.3402	0.0675	1.0837	1.3483
TRANSPORTE	1.0416	0.0613	0.9215	1.1617	0.0663	0.9116	1.1716
EDUCACION	0.8639	0.1225	0.6238	1.1039	0.1499	0.5700	1.1577
OTROS	1.1684	0.0924	0.9874	1.3495	0.0961	0.9802	1.3567
INDICE KISH	**1.0156**	**0.0234**	**0.9697**	**1.0615**	**0.0127**	**0.9907**	**1.0406**

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{R}_4$	$se\left(\hat{R}_4\right)$	**INF**	**SUP**	$se\left(\hat{R}_4\right)$	**INF**	**SUP**
ALIMENTOS	0.9896	0.0582	0.8756	0.9896	0.0585	0.8749	1.1037
ROPA	0.9824	0.0514	0.8816	0.9824	0.0524	0.8798	1.0832
VIVIENDA	0.9600	0.0415	0.8786	0.9600	0.0429	0.8760	1.0414
MUEBLES	0.5896	0.0403	0.5105	0.5896	0.0422	0.5068	0.6686
SALUD	1.2009	0.0631	1.0772	1.2009	0.0667	1.0702	1.3246
TRANSPORTE	1.0110	0.0488	0.9153	1.0110	0.0539	0.9054	1.1067
EDUCACION	0.8429	0.1225	0.6029	0.8429	0.1451	0.5586	1.0830
OTROS	2.0363	0.1590	1.7247	2.0363	0.1635	1.7159	2.3478
INDICE KISH	**1.0766**	**0.0280**	**1.0217**	**1.0766**	**0.0197**	**1.0381**	**1.1315**

Tabla C.4: Estimación de Razón del Gasto Mensual Promedio a nivel **Urbano** para cada una de las categorías respecto al Gasto Total, para los años 2000, $\hat{r}_0$, 2002, $\hat{r}_2$, y 2004, $\hat{r}_4$. Estimación de su desviación estándar, $se\left(\hat{r}_0\right)$, $se\left(\hat{r}_2\right)$ y $se\left(\hat{r}_4\right)$, y límites de confianza al 95 % de la estimación.

CATEGORIA	$\hat{r}_0$	$se\left(\hat{r}_0\right)$	**INF**	**SUP**
ALIMENTOS	23.68	1.27	21.20	26.16
ROPA	6.71	0.33	6.07	7.36
VIVIENDA	17.44	0.68	16.11	18.78
MUEBLES	8.30	0.54	7.25	9.35
SALUD	6.33	0.32	5.71	6.95
TRANSPORTE	15.47	0.72	14.06	16.88
EDUCACION	16.62	2.30	12.12	21.13
OTROS	5.44	0.40	4.65	6.22

CATEGORIA	$\hat{r}_2$	$se\left(\hat{r}_2\right)$	**INF**	**SUP**
ALIMENTOS	24.88	0.64	23.63	26.13
ROPA	7.08	0.16	6.76	7.39
VIVIENDA	17.47	0.37	16.75	18.19
MUEBLES	6.05	0.18	5.70	6.40
SALUD	7.70	0.16	7.38	8.01
TRANSPORTE	16.11	0.64	14.86	17.36
EDUCACION	14.36	0.55	13.28	15.44
OTROS	6.36	0.21	5.94	6.77

CATEGORIA	$\hat{r}_4$	$se\left(\hat{r}_4\right)$	**INF**	**SUP**
ALIMENTOS	23.44	0.52	22.41	24.46
ROPA	6.60	0.11	6.39	6.80
VIVIENDA	16.74	0.31	16.14	17.34
MUEBLES	4.89	0.10	4.69	5.10
SALUD	7.60	0.16	7.28	7.92
TRANSPORTE	15.64	0.32	15.02	16.26
EDUCACION	14.01	0.47	13.10	14.93
OTROS	11.08	0.32	10.45	11.70

Tabla C.5: Estimación del error estándar, $se\left(\hat{r}'_2\right)$ y $se\left(\hat{r}'_4\right)$, para los estimadores de razón, $\hat{r}'_2\left(=\frac{\hat{t}_{A02}}{\hat{t}_{A00}}\right)$ y $\hat{r}'_4\left(=\frac{\hat{t}_{A04}}{\hat{t}_{A00}}\right)$, a nivel **Urbano** obtenidos mediante el método propuesto por: **KISH.** (1968) y **JK.** la técnica jackknife aplicada al estimador de razón. Además, cociente de la estimación del error estándar de los métodos mencionados para los años 2002 y 2004, respectivamente.

	2002/2000			**2004/2000**		
CATEGORIA	**KISH**	**JK**	**JK/KISH**	**KISH**	**JK**	**JK/KISH**
ALIMENTOS	13.04	16.36	1.25	15.28	19.23	1.26
ROPA	14.71	18.15	1.23	16.98	21.21	1.25
VIVIENDA	14.14	18.15	1.28	17.37	21.56	1.24
MUEBLES	9.43	11.57	1.23	9.40	11.57	1.23
SALUD	15.50	20.03	1.29	20.45	25.01	1.22
TRANSPORTE	18.55	21.97	1.18	21.75	26.15	1.20
EDUCACION	21.99	27.69	1.26	27.87	33.69	1.21
OTROS	17.91	22.74	1.27	39.25	48.07	1.22
GASTO TOTAL	14.41	17.89	1.24	18.23	22.24	1.22

Tabla C.6: Estimación del error estándar, $se\left(\hat{R}_2\right)$ y $se\left(\hat{R}_4\right)$, para los estimadores de las relativas, $\hat{R}_2\left(=\frac{\hat{r}_2}{\hat{r}_0}\right)$, y $\hat{R}_4\left(=\frac{\hat{r}_4}{\hat{r}_0}\right)$, y del índice de Kish, $\hat{I}_2\left(=\frac{1}{J}\sum_{j=1}^{J}\hat{R}_{j2}\right)$ e $\hat{I}_4\left(=\frac{1}{J}\sum_{j=1}^{J}\hat{R}_{j4}\right)$, a nivel **Urbano** obtenidos mediante el método propuesto por: **KISH.** (1968) y **JK.** la técnica jackknife aplicada al estimador de las relativas y al índice de Kish. Además, cociente de la estimación del error estándar de los métodos mencionados para los años 2002 y 2004, respectivamente.

	2002/2000			**2004/2000**		
CATEGORIA	**KISH**	**JK**	**JK/KISH**	**KISH**	**JK**	**JK/KISH**
ALIMENTOS	0.0572	0.0638	1.12	0.0582	0.0585	1.01
ROPA	0.0602	0.0588	0.98	0.0514	0.0524	1.02
VIVIENDA	0.0434	0.0458	1.06	0.0415	0.0429	1.03
MUEBLES	0.0519	0.0545	1.05	0.0403	0.0422	1.05
SALUD	0.0634	0.0675	1.06	0.0631	0.0667	1.06
TRANSPORTE	0.0613	0.0663	1.08	0.0488	0.0539	1.10
EDUCACION	0.1225	0.1499	1.22	0.1225	0.1451	1.18
OTROS	0.0924	0.0961	1.04	0.1590	0.1635	1.03
INDICE KISH	**0.0234**	**0.0127**	**0.54**	**0.0280**	**0.0197**	**0.70**

C.1.2. Estimaciones a Nivel Rural

Tabla C.7: Estimación del Gasto Mensual Promedio en Hogares Mexicanos a nivel **Rural** por categoría, en los años 2000, 2002 y 2004. Información obtenida de la ENIGH 2000, 2002 y 2004. Las categorías son las consideradas en el INPC.

CATEGORIA	**GASTO TOTAL DE LOS AÑOS**		
	2000	**2002**	**2004**
ALIMENTOS	4,862,432,566	5,720,129,400	7,923,644,208
ROPA	950,983,077	1,320,676,622	1,463,704,502
VIVIENDA	1,457,479,448	2,182,340,676	2,721,079,235
MUEBLES	1,506,554,070	1,228,621,030	1,449,665,848
SALUD	1,088,283,015	1,664,706,992	2,169,385,068
TRANSPORTE	1,614,392,684	2,869,915,074	3,217,874,968
EDUCACION	814,729,400	1,443,334,233	1,463,011,915
OTROS	420,667,310	1,504,551,149	2,165,556,651
GASTO TOTAL	12,715,521,570	17,934,275,176	22,573,922,396

Tabla C.8: Estimación de Razón del Gasto Mensual Promedio a nivel **Rural** por categoría, para los años 2002, $\hat{r}'_2 \left(= \frac{\hat{t}_{A02}}{\hat{t}_{A00}}\right)$, y 2004, $\hat{r}'_4 \left(= \frac{\hat{t}_{A04}}{\hat{t}_{A00}}\right)$, así como estimación de su desviación estándar, $se\left(\hat{r}'_2\right)$ y $se\left(\hat{r}'_4\right)$, y límites de confianza de la estimación al 95 %, considerando el método propuesto por Kish (1968) y empleando la técnica jackknife, tomando como base de comparación el año 2000.

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{r}'_2$	$se\left(\hat{r}'_2\right)$	**INF**	**SUP**	$se\left(\hat{r}'_2\right)$	**INF**	**SUP**
ALIMENTOS	117.64	13.24	91.69	143.59	15.80	86.66	148.62
ROPA	138.87	20.70	98.30	179.45	23.18	93.45	184.30
VIVIENDA	149.73	46.54	58.52	240.95	48.14	55.37	244.10
MUEBLES	81.55	10.52	60.94	102.17	12.14	57.76	105.34
SALUD	152.97	20.59	112.60	193.33	23.96	106.01	199.92
TRANSPORTE	177.77	56.82	66.40	289.14	58.73	62.66	292.88
EDUCACION	177.16	66.26	47.28	307.03	67.70	44.47	309.84
OTROS	357.66	229.71	0.00	807.89	231.96	0.00	812.30
GASTO TOTAL	141.04	29.60	83.03	199.05	31.37	79.57	202.52

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{r}'_4$	$se\left(\hat{r}'_4\right)$	**INF**	**SUP**	$se\left(\hat{r}'_4\right)$	**INF**	**SUP**
ALIMENTOS	162.96	20.50	122.78	203.13	23.88	116.16	209.76
ROPA	153.91	19.62	115.46	192.37	22.72	109.38	198.45
VIVIENDA	186.70	28.21	131.40	241.99	32.00	123.98	249.41
MUEBLES	96.22	13.01	70.73	121.72	15.09	66.65	125.79
SALUD	199.34	29.11	142.29	256.39	33.29	134.08	264.60
TRANSPORTE	199.32	35.41	129.91	268.73	39.46	121.98	276.67
EDUCACION	179.57	26.89	126.86	232.28	30.32	120.15	238.99
OTROS	514.79	85.16	347.88	681.70	96.91	324.85	704.73
GASTO TOTAL	177.53	22.77	132.90	222.16	26.44	125.72	229.35

Tabla C.9: Estimación de las Relativas del Gasto Mensual Promedio a nivel **Rural** para cada una de las categorías consideradas en el INPC, y estimación del índice de Kish para los años 2002, $\hat{R}_2 \left(= \frac{\hat{r}_2}{\hat{r}_0}\right)$, y 2004, $\hat{R}_4 \left(= \frac{\hat{r}_4}{\hat{r}_0}\right)$, así como estimación de su desviación estándar, $se\left(\hat{R}_2\right)$ y $se\left(\hat{R}_4\right)$, y límites de confianza al 95 % de la estimación, considerando el método propuesto por Kish (1968) y empleando la técnica jackknife, tomando como base de comparación el año 2000.

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{R}_2$	$se\left(\hat{R}_2\right)$	**INF**	**SUP**	$se\left(\hat{R}_2\right)$	**INF**	**SUP**
ALIMENTOS	0.8341	0.1407	0.5584	1.1098	0.1721	0.4967	1.1714
ROPA	0.9846	0.0911	0.8062	1.1631	0.1092	0.7706	1.1987
VIVIENDA	1.0616	0.1271	0.8126	1.3107	0.1549	0.7581	1.3652
MUEBLES	0.5782	0.0768	0.4277	0.7287	0.0928	0.3964	0.7600
SALUD	1.0845	0.1808	0.7301	1.4390	0.2167	0.6597	1.5093
TRANSPORTE	1.2604	0.1643	0.9383	1.5825	0.1937	0.8807	1.6401
EDUCACION	1.2560	0.2331	0.7991	1.7130	0.2815	0.7043	1.8078
OTROS	2.5358	1.1595	0.2632	4.8085	1.4142	0.0000	5.3077
INDICE KISH	**1.1994**	**0.1530**	**0.8996**	**1.4992**	**0.1785**	**0.8496**	**1.5492**

		KISH			**JACKKNIFE**		
CATEGORIA	$\hat{R}_4$	$se\left(\hat{R}_4\right)$	**INF**	**SUP**	$se\left(\hat{R}_4\right)$	**INF**	**SUP**
ALIMENTOS	0.9179	0.0314	0.8564	0.9179	0.0344	0.8505	0.9794
ROPA	0.8670	0.0393	0.7899	0.8670	0.0413	0.7860	0.9441
VIVIENDA	1.0516	0.0695	0.9155	1.0516	0.0740	0.9066	1.1878
MUEBLES	0.5420	0.0255	0.4920	0.5420	0.0272	0.4888	0.5920
SALUD	1.1229	0.0709	0.9839	1.1229	0.0749	0.9760	1.2618
TRANSPORTE	1.1228	0.1088	0.9094	1.1228	0.1153	0.8968	1.3361
EDUCACION	1.0115	0.0854	0.8442	1.0115	0.0884	0.8383	1.1788
OTROS	2.8997	0.3193	2.2740	2.8997	0.3431	2.2272	3.5255
INDICE KISH	**1.1919**	**0.0418**	**1.1100**	**1.1919**	**0.0408**	**1.1119**	**1.2738**

Tabla C.10: Estimación de Razón del Gasto Mensual Promedio a nivel **Rural** para cada una de las categorías respecto al Gasto Total, para los años 2000, $\hat{r}_0$, 2002, $\hat{r}_2$, y 2004, $\hat{r}_4$. Estimación de su desviación estándar, $se(\hat{r}_0)$, $se(\hat{r}_2)$ y $se(\hat{r}_4)$, y límites de confianza al 95 % de la estimación.

CATEGORIA	$\hat{r}_0$	$se(\hat{r}_0)$	**INF**	**SUP**
ALIMENTOS	38.24	0.99	36.30	40.18
ROPA	7.48	0.24	7.01	7.94
VIVIENDA	11.46	0.48	10.53	12.39
MUEBLES	11.85	0.35	11.16	12.53
SALUD	8.56	0.47	7.64	9.48
TRANSPORTE	12.70	0.87	11.00	14.39
EDUCACION	6.41	0.33	5.76	7.05
OTROS	3.31	0.32	2.68	3.93

CATEGORIA	$\hat{r}_2$	$se(\hat{r}_2)$	**INF**	**SUP**
ALIMENTOS	31.89	5.34	21.43	42.36
ROPA	7.36	0.64	6.10	8.63
VIVIENDA	12.17	1.39	9.44	14.90
MUEBLES	6.85	0.89	5.11	8.59
SALUD	9.28	1.46	6.41	12.15
TRANSPORTE	16.00	1.78	12.50	19.50
EDUCACION	8.05	1.44	5.22	10.88
OTROS	8.39	3.76	1.02	15.76

CATEGORIA	$\hat{r}_4$	$se(\hat{r}_4)$	**INF**	**SUP**
ALIMENTOS	35.10	0.86	33.42	36.78
ROPA	6.48	0.22	6.05	6.91
VIVIENDA	12.05	0.64	10.80	13.31
MUEBLES	6.42	0.25	5.93	6.91
SALUD	9.61	0.32	8.98	10.24
TRANSPORTE	14.25	1.00	12.29	16.22
EDUCACION	6.48	0.44	5.63	7.34
OTROS	9.59	0.52	8.58	10.61

Tabla C.11: Estimación del error estándar, $se\left(\hat{r}'_2\right)$ y $se\left(\hat{r}'_4\right)$, para los estimadores de razón, $\hat{r}'_2\left(=\frac{\hat{t}_{A02}}{\hat{t}_{A00}}\right)$ y $\hat{r}'_4\left(=\frac{\hat{t}_{A04}}{\hat{t}_{A00}}\right)$, a nivel **Rural** obtenidos mediante el método propuesto por: **KISH.** (1968) y **JK.** la técnica jackknife aplicada al estimador de razón. Además, cociente de la estimación del error estándar de los métodos mencionados para los años 2002 y 2004, respectivamente.

	2002/2000			**2004/2000**		
CATEGORIA	**KISH**	**JK**	**JK/KISH**	**KISH**	**JK**	**JK/KISH**
ALIMENTOS	13.24	15.80	1.19	20.50	23.88	1.16
ROPA	20.70	23.18	1.12	19.62	22.72	1.16
VIVIENDA	46.54	48.14	1.03	28.21	32.00	1.13
MUEBLES	10.52	12.14	1.15	13.01	15.09	1.16
SALUD	20.59	23.96	1.16	29.11	33.29	1.14
TRANSPORTE	56.82	58.73	1.03	35.41	39.46	1.11
EDUCACION	66.26	67.70	1.02	26.89	30.32	1.13
OTROS	229.71	231.96	1.01	85.16	96.91	1.14
GASTO TOTAL	29.60	31.37	1.06	22.77	26.44	1.16

Tabla C.12: Estimación del error estándar, $se\left(\hat{R}_2\right)$ y $se\left(\hat{R}_4\right)$, para los estimadores de las relativas, $\hat{R}_2\left(=\frac{\hat{r}_2}{\hat{r}_0}\right)$, y $\hat{R}_4\left(=\frac{\hat{r}_4}{\hat{r}_0}\right)$, y del índice de Kish, $\hat{I}_2\left(=\frac{1}{J}\sum_{j=1}^{J}\hat{R}_{j2}\right)$ e $\hat{I}_4\left(=\frac{1}{J}\sum_{j=1}^{J}\hat{R}_{j4}\right)$, a nivel **Rural** obtenidos mediante el método propuesto por: **KISH.** (1968) y **JK.** la técnica jackknife aplicada al estimador de las relativas y al índice de Kish. Además, cociente de la estimación del error estándar de los métodos mencionados para los años 2002 y 2004, respectivamente.

	2002/2000			**2004/2000**		
CATEGORIA	**KISH**	**JK**	**JK/KISH**	**KISH**	**JK**	**JK/KISH**
ALIMENTOS	0.1407	0.1721	1.22	0.0314	0.0344	1.09
ROPA	0.0911	0.1092	1.20	0.0393	0.0413	1.05
VIVIENDA	0.1271	0.1549	1.22	0.0695	0.0740	1.07
MUEBLES	0.0768	0.0928	1.21	0.0255	0.0272	1.06
SALUD	0.1808	0.2167	1.20	0.0709	0.0749	1.06
TRANSPORTE	0.1643	0.1937	1.18	0.1088	0.1153	1.06
EDUCACION	0.2331	0.2815	1.21	0.0854	0.0884	1.04
OTROS	1.1595	1.4142	1.22	0.3193	0.3431	1.07
INDICE KISH	**0.1530**	**0.1785**	**1.17**	**0.0418**	**0.0408**	**0.98**

C.1.3. Estimaciones para la Zona Metropolitana de la Ciudad de México

Tabla C.13: Estimación del Gasto Mensual Promedio en Hogares Mexicanos a nivel **Zona Metropolitana de la Ciudad de México** por categoría, en los años 2000, 2002 y 2004. Información obtenida de la ENIGH 2000, 2002 y 2004. Las categorías son las consideradas en el INPC.

CATEGORIA	**GASTO TOTAL DE LOS AÑOS**		
	2000	**2002**	**2004**
ALIMENTOS	6,515,922,814	8,435,020,724	9,337,698,044
ROPA	2,137,305,618	2,324,565,878	2,753,149,259
VIVIENDA	5,673,480,566	6,098,666,750	7,151,139,634
MUEBLES	1,856,336,400	1,835,904,476	1,678,000,615
SALUD	1,766,988,221	2,348,302,693	3,064,104,772
TRANSPORTE	4,689,053,996	4,963,913,270	6,377,134,365
EDUCACION	5,275,389,532	5,736,356,104	6,349,892,221
OTROS	1,941,978,448	2,576,233,757	5,244,361,588
GASTO TOTAL	29,856,455,596	34,318,963,652	41,955,480,498

Tabla C.14: Estimación de Razón del Gasto Mensual Promedio a nivel **Zona Metropolitana de la Ciudad de México** por categoría, para los años 2002, $\hat{r}'_2 \left(= \frac{\hat{t}_{A02}}{\hat{t}_{A00}}\right)$, y 2004, $\hat{r}'_4 \left(= \frac{\hat{t}_{A04}}{\hat{t}_{A00}}\right)$, así como estimación de su desviación estándar, $se\left(\hat{r}'_2\right)$ y $se\left(\hat{r}'_4\right)$, y límites de confianza de la estimación al 95 %, empleando la técnica jackknife, tomando como base de comparación el año 2000.

CATEGORIA	$\hat{r}'_2$	$se\left(\hat{r}'_2\right)$	**INF**	**SUP**
ALIMENTOS	129.45	50.73	30.03	228.87
ROPA	108.76	46.88	16.89	200.64
VIVIENDA	107.49	46.18	16.98	198.01
MUEBLES	98.90	40.56	19.40	178.39
SALUD	132.90	55.85	23.42	242.37
TRANSPORTE	105.86	47.38	13.01	198.72
EDUCACION	108.74	55.27	0.41	217.07
OTROS	132.66	54.62	25.60	239.72
GASTO TOTAL	114.95	47.88	21.10	208.79

CATEGORIA	$\hat{r}'_4$	$se\left(\hat{r}'_4\right)$	**INF**	**SUP**
ALIMENTOS	143.31	54.38	36.72	249.89
ROPA	128.81	54.13	22.72	234.91
VIVIENDA	126.05	52.93	22.30	229.79
MUEBLES	90.39	35.41	20.99	159.80
SALUD	173.41	71.26	33.74	313.07
TRANSPORTE	136.00	59.17	20.03	251.97
EDUCACION	120.37	59.52	3.70	237.03
OTROS	270.05	108.64	57.13	482.98
GASTO TOTAL	140.52	56.79	29.22	251.83

Tabla C.15: Estimación de Razón del Gasto Mensual Promedio a nivel **Zona Metropolitana de la Ciudad de México** para cada una de las categorías respecto al Gasto Total, para los años 2000, $\hat{r}_0$, 2002, $\hat{r}_2$, y 2004, $\hat{r}_4$. Estimación de su desviación estándar, $se\,(\hat{r}_0)$, $se\,(\hat{r}_2)$ y $se\,(\hat{r}_4)$, y límites de confianza al 95 % de la estimación.

CATEGORIA	$\hat{r}_0$	$se\,(\hat{r}_0)$	**INF**	**SUP**
ALIMENTOS	21.82	1.80	18.30	25.35
ROPA	7.16	0.37	6.43	7.88
VIVIENDA	19.00	1.32	16.41	21.60
MUEBLES	6.22	0.85	4.54	7.89
SALUD	5.92	0.20	5.52	6.31
TRANSPORTE	15.71	1.18	13.40	18.01
EDUCACION	17.67	1.77	14.20	21.14
OTROS	6.50	0.56	5.41	7.60

CATEGORIA	$\hat{r}_2$	$se\,(\hat{r}_2)$	**INF**	**SUP**
ALIMENTOS	24.58	1.47	21.70	27.46
ROPA	6.77	0.25	6.28	7.26
VIVIENDA	17.77	0.87	16.06	19.48
MUEBLES	5.35	0.49	4.39	6.31
SALUD	6.84	0.28	6.30	7.38
TRANSPORTE	14.46	0.47	13.54	15.39
EDUCACION	16.71	1.37	14.04	19.39
OTROS	7.51	0.39	6.75	8.26

CATEGORIA	$\hat{r}_4$	$se\,(\hat{r}_4)$	**INF**	**SUP**
ALIMENTOS	22.26	1.43	19.45	25.07
ROPA	6.56	0.22	6.14	6.99
VIVIENDA	17.04	0.76	15.56	18.53
MUEBLES	4.00	0.21	3.59	4.41
SALUD	7.30	0.30	6.72	7.89
TRANSPORTE	15.20	0.72	13.79	16.61
EDUCACION	15.13	1.13	12.91	17.35
OTROS	12.50	0.90	10.73	14.27

C.2. Resultados Gráficos

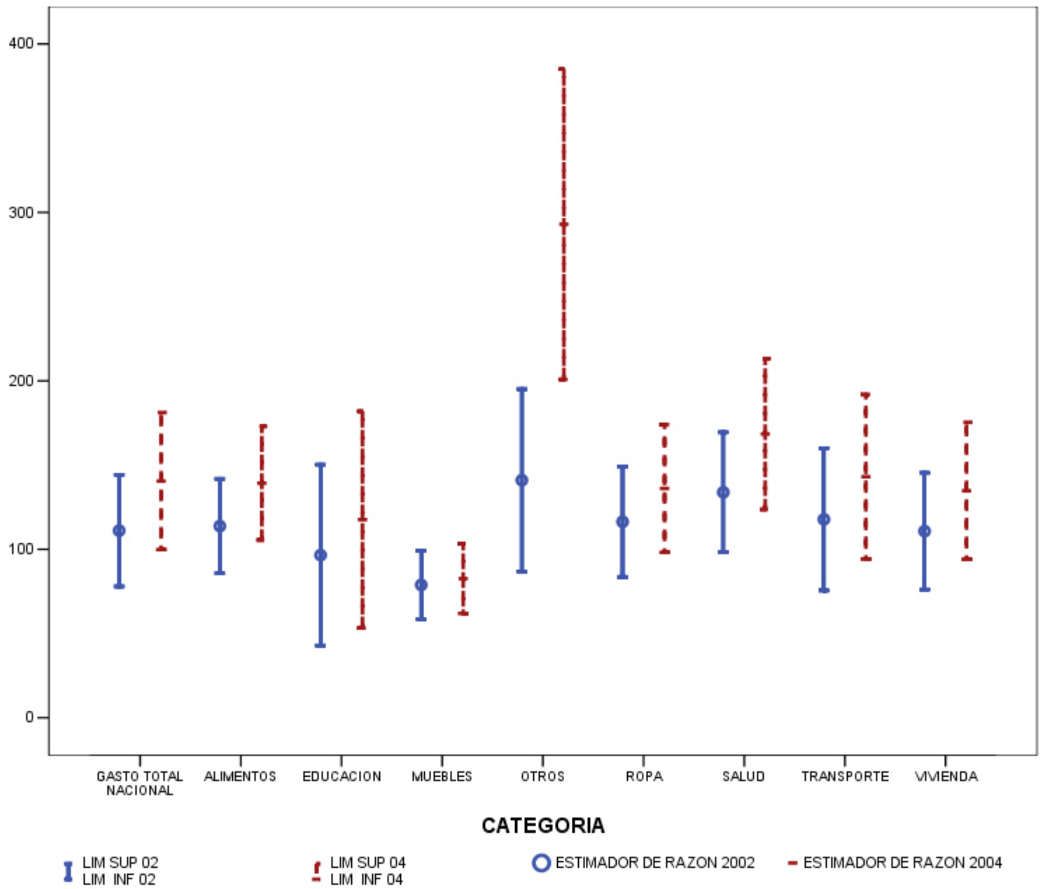

Figura C.1: Estimación de Razón y de su intervalo de confianza al 95 % a nivel **Nacional** para los años 2002 y 2004 respecto al año 2000 para las categorías consideradas. La estimación de la varianza para estos intervalos se obtuvo mediante el método propuesto por Kish.

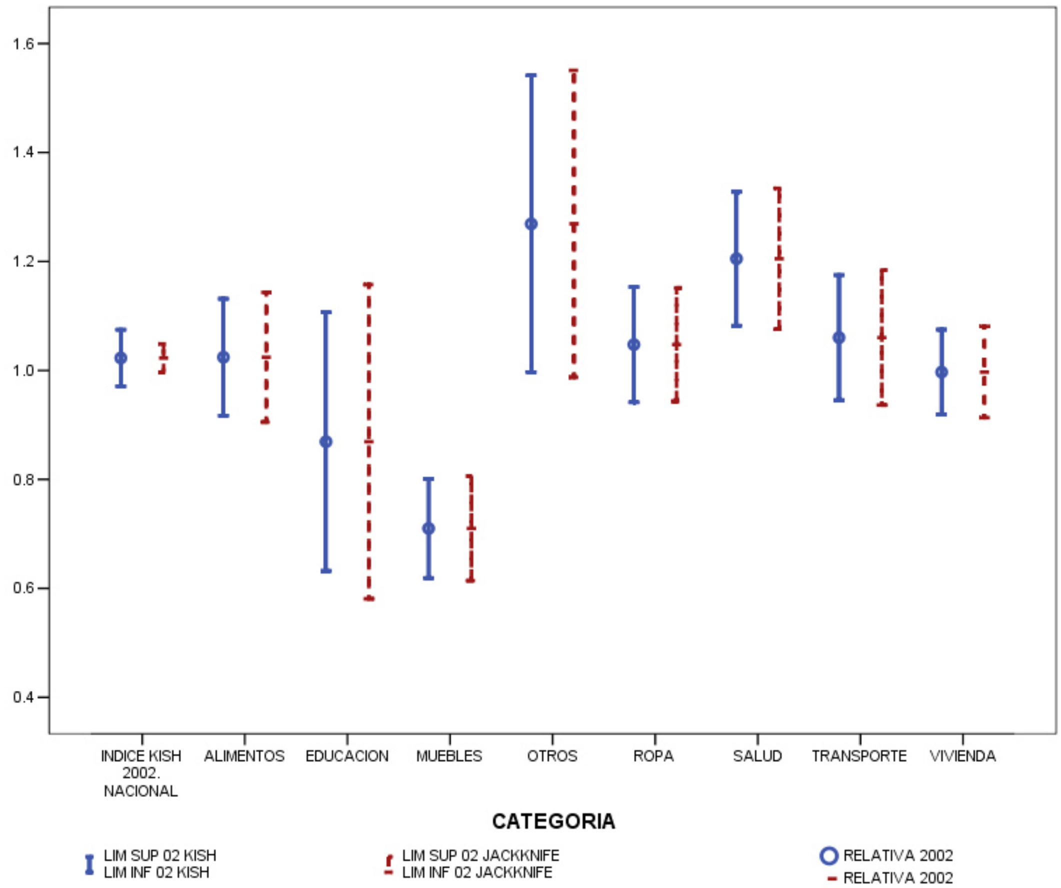

Figura C.2: Estimador de Indice de Kish, de sus relativas por categoría, y de su intervalo de confianza al 95 % a nivel **Nacional** para el año 2002 respecto al año 2000. La estimación de la varianza para estos intervalos se obtuvo mediante el método propuesto por Kish.

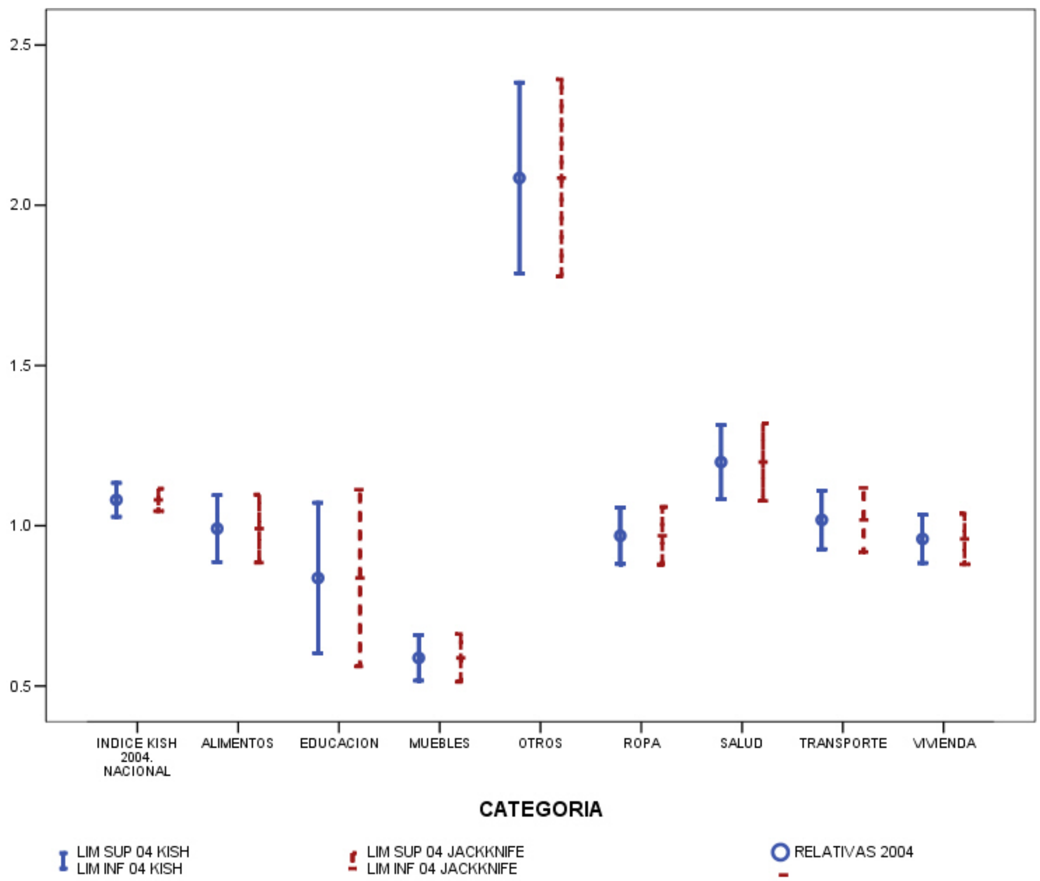

Figura C.3: Estimador de Indice de Kish, de sus relativas por categoría, y de su intervalo de confianza al 95 % a nivel **Nacional** para el año 2004 respecto al año 2000. La estimación de la varianza para estos intervalos se obtuvo mediante el método propuesto por Kish.

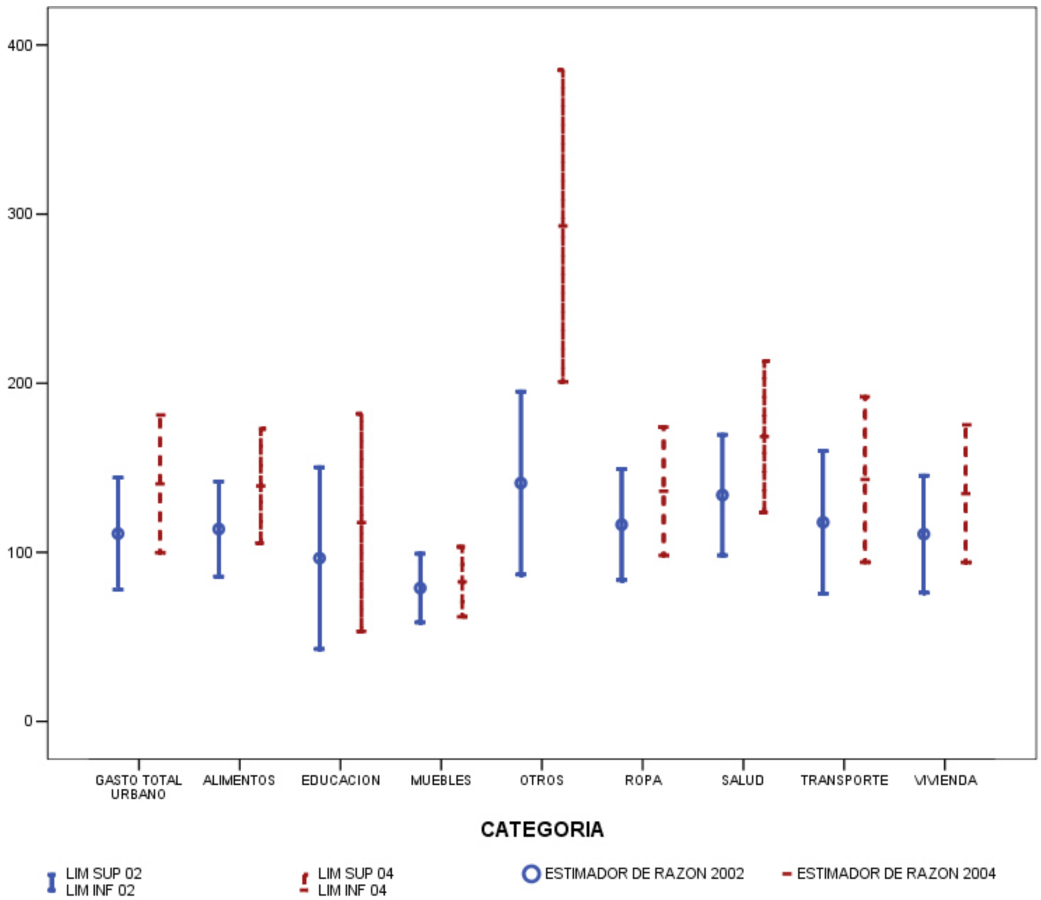

Figura C.4: Estimación de Razón y de su intervalo de confianza al 95 % a nivel **Urbano** para los años 2002 y 2004 respecto al año 2000 para las categorías consideradas. La estimación de la varianza para estos intervalos se obtuvo mediante el método propuesto por Kish.

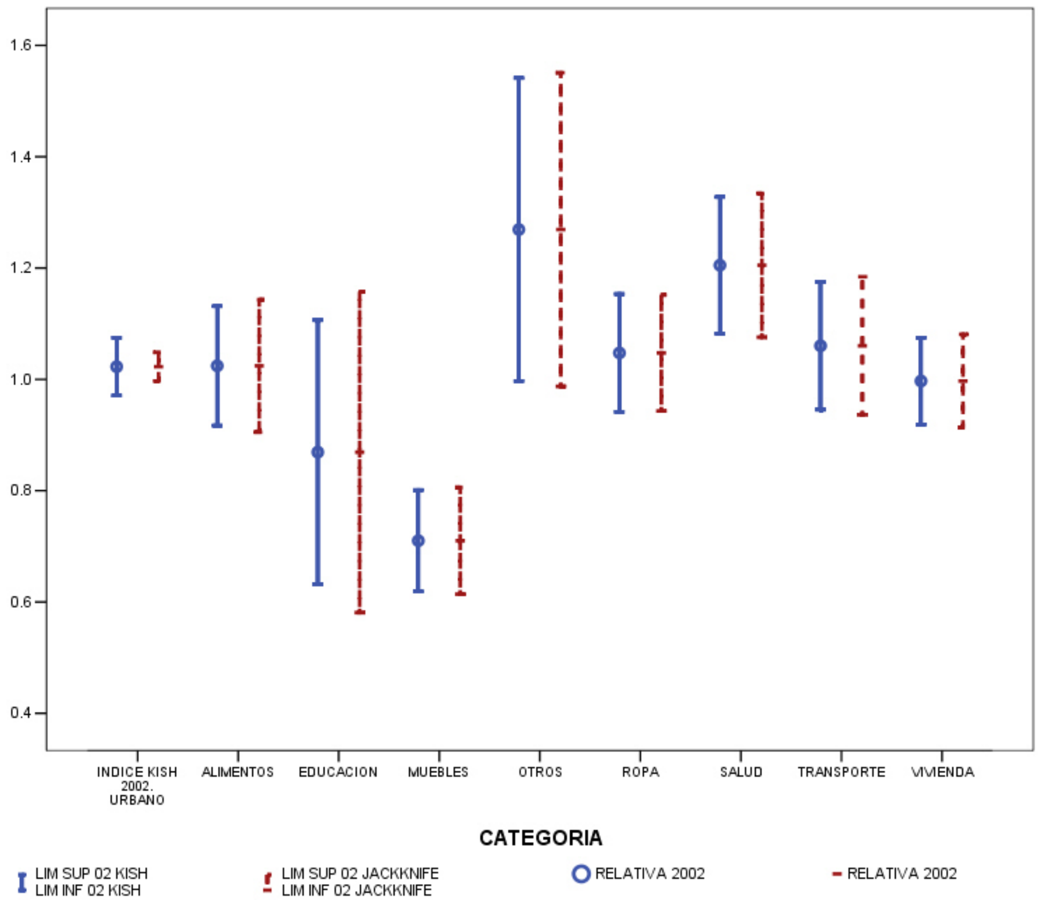

Figura C.5: Estimador de Indice de Kish, de sus relativas por categoría, y de su intervalo de confianza al 95 % a nivel **Urbano** para el año 2002 respecto al año 2000. La estimación de la varianza para estos intervalos se obtuvo mediante el método propuesto por Kish.

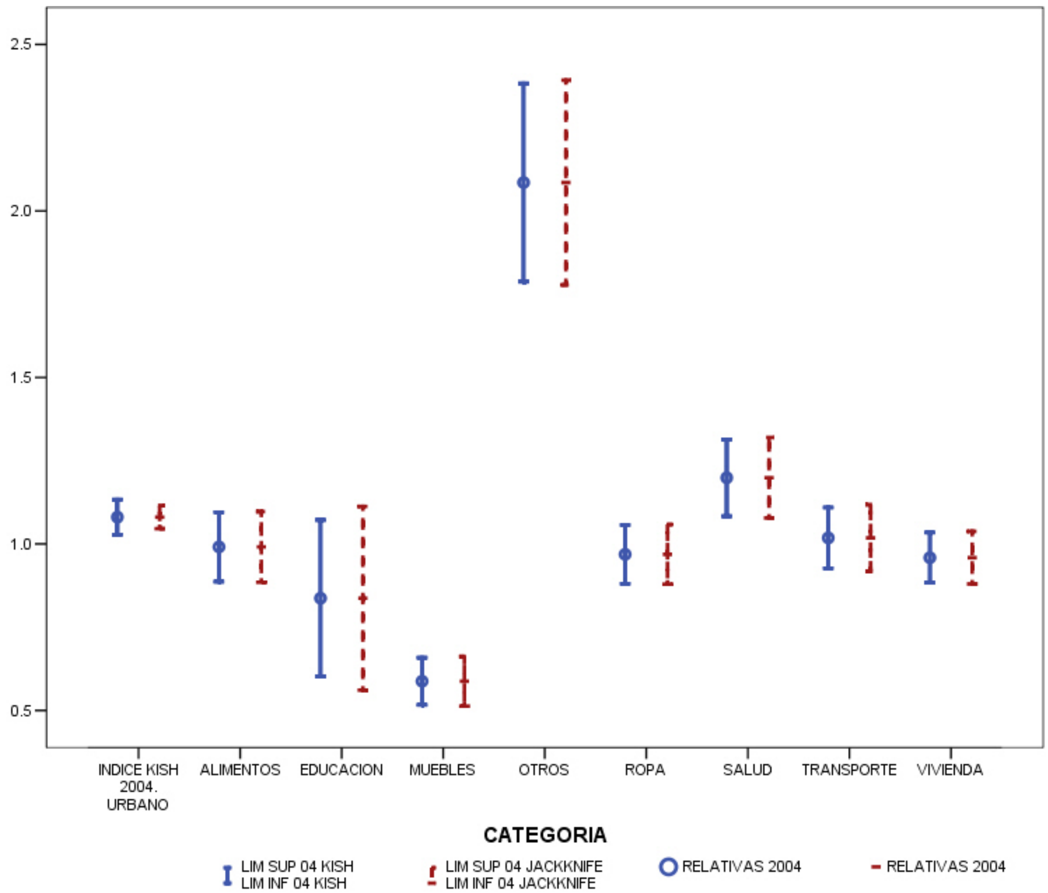

Figura C.6: Estimador de Indice de Kish, de sus relativas por categoría, y de su intervalo de confianza al 95 % a nivel **Urbano** para el año 2004 respecto al año 2000. La estimación de la varianza para estos intervalos se obtuvo mediante el método propuesto por Kish.

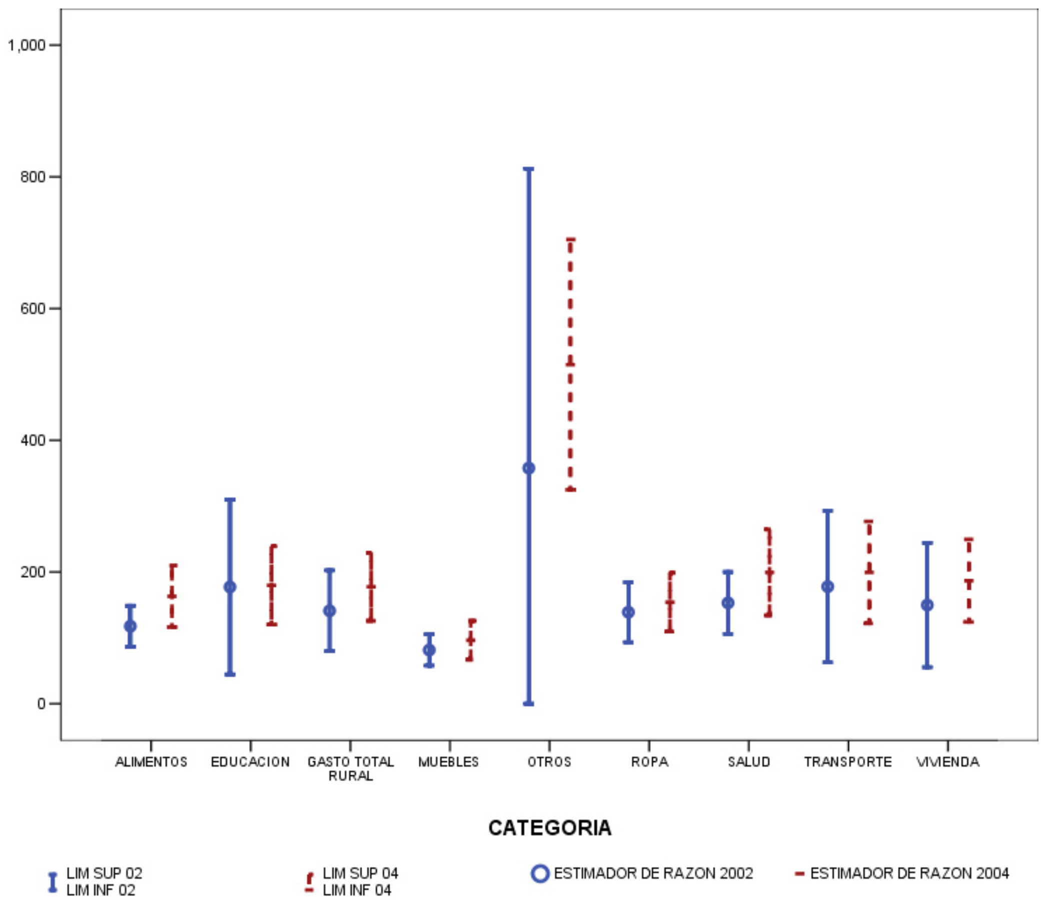

Figura C.7: Estimación de Razón y de su intervalo de confianza al 95 % a nivel **Rural** para los años 2002 y 2004 respecto al año 2000 para las categorías consideradas. La estimación de la varianza para estos intervalos se obtuvo mediante el método propuesto por Kish.

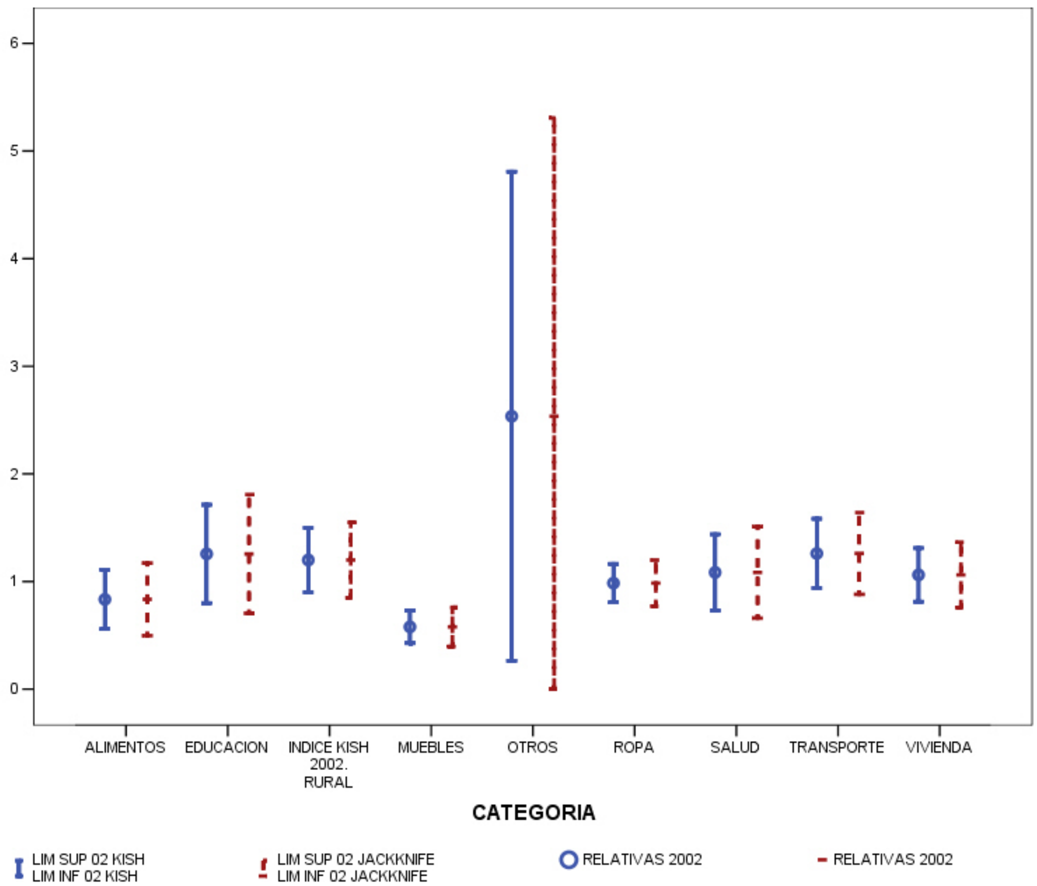

Figura C.8: Estimador de Indice de Kish, de sus relativas por categoría, y de su intervalo de confianza al 95 % a nivel **Rural** para el año 2002 respecto al año 2000. La estimación de la varianza para estos intervalos se obtuvo mediante el método propuesto por Kish.

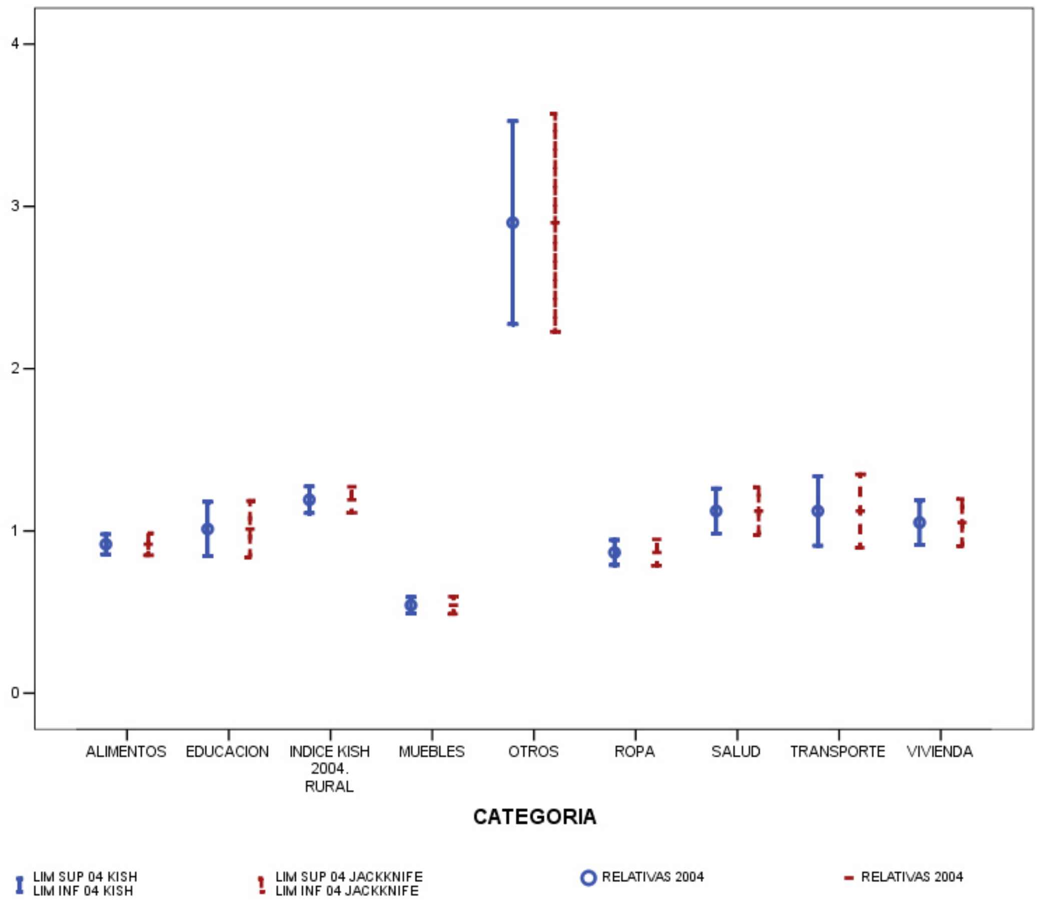

Figura C.9: Estimador de Indice de Kish, de sus relativas por categoría, y de su intervalo de confianza al 95 % a nivel **Rural** para el año 2004 respecto al año 2000. La estimación de la varianza para estos intervalos se obtuvo mediante el método propuesto por Kish.

Bibliografía

[1] Banco de México, Junio, 2002. El Indice Nacional de Precios al Consumidor: Características y Actualización de su Base al Año 2002.

[2] Cochran, W. G. (1976). Técnicas de Muestreo. 6a Impresión. Wiley & Sons, Inc.

[3] Consejo Nacional de Población (2003). La delimitación de zonas metropolitanas. Primera edición.,

[4] Contar 2000, Sistema para la consulta de tabulados y bases de datos de la muestra, INEGI.

[5] XII Censo General de Población y Vivienda del 2000. México.

`http://www.inegi.gob.mx`

[6] Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika* **46**, 477-480.

[7] Fisher, R. A. (1921). On the Mathematical Foundations of Theoretical Statistics. *Reproduced from the Philosophical Transactions of the Royal Society, A.* **222**, 309-368 (1922).

[8] Francisco, C. A. and Fuller, W. A. (1991). Quantile Estimation with a Complex Survey Desing. *Annals of Statistics.* **19**, 454-469.

[9] Kendall, M., and Stuart, A. (1976). The Advanced Theory of Statistics. Vol. 1, Distribution Theory. 4th Ed. Charles Griffin & Company Limited. London.

[10] Kish, L. (1968). Standard Errors for Indexes from Complex Samples. *Journal of the American Statistical Association.* Vol. 63, No. 322.

[11] Kish, L. (1979). Muestreo de Encuestas. Segunda reimpresión. Ed. Trillas.

[12] Research Triangle Institute (2001). SUDAAN User's Manual, Release 8.0

`www.rti.org/sudaan`

[13] Medina, F. (2001). Consideraciones sobre el índice de Gini para medir la concentración del ingreso. Serie: Estudios Estadísticos y Prospectivos. Publicación de las Naciones Unidas. Santiago de Chile.

[14] Nygård, F., and Sandström, A. (1985). Estimation of the Gini and the enthropy inequality parameters in finite populations. *Journal of Official Statistics.* **1**, 399-412.

[15] Ogwang, T. (2000). A convenient method of computing the Gini index and its standard error. *Oxford Bulletin of Economics and Statistics.* **62**, 1, 123-129.

[16] Página del Banco de México.

`http://www.banxico.org.mx`

[17] Página donde se puede descargar el paquete **R**.

`http://cran.r-project.org`

[18] Quenouille, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics.* **20**, 355-375.

[19] Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika.* **43**, 353-360.

[20] Rao, J. N. K., Kovar, J. G., and Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika.* **77**, 365-375.

[21] Sandström, A., Wretman, J. H., and Waldén, B. (1988). Variance estimators of the Gini Coefficient-Probability Sampling. *Journal of Business & Economic Statistics.* **6**, 113-119.

[22] Särndal, C. E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York. Springer-Verlag.

[23] Síntesis Metodológica de la Encuesta Nacional de Ingresos y Gastos en los Hogares. ENIGH 2000. INEGI.

`http://www.inegi.gob.mx`

[24] Síntesis Metodológica de la Encuesta Nacional de Ingresos y Gastos en los Hogares. ENIGH 2002. INEGI.

`http://www.inegi.gob.mx`

[25] Síntesis Metodológica de la Encuesta Nacional de Ingresos y Gastos en los Hogares. ENIGH 2004. INEGI.

`http://www.inegi.gob.mx`

[26] Stuart, Alan & Ord, Keith, (1994). Kendall´s Advanced Theory of Statistics, Sixth Edition, Volume I, Distribution Theory. Arnold Publishers.

[27] Tukey, J. W. (1958). Bias and confidence in not-quite large samples (abstract). *Annals of Mathematical Statistics* **29**, 614.

[28] Wolter, K. M. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

[29] Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association.* **47**, 635-646.

[30] Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association.* **66**, 411-414.